Chapter 27:

Crime Trip Generation

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Chapter 27:
Crime Trip Generation

Background

In this chapter, the theory and mechanics of the trip generation stage will be explained. *Trip generation* is a model of the number of trips that originate and end in each zone for a given jurisdiction. Given a set of *N* destination zones and *M* origin zones (which include all the destination zones and, possibly, zones from adjacent jurisdictions), separate models are produced of the number of crimes originating and ending in each of these zones. That is, a separate model is produced of the number of crimes originating in each of the *M* origin zones, and another model is produced of the number of crimes ending in each of the *N* destination zones. The first is a *crime production* model while the second is a *crime attraction* model.

Two points should be emphasized. First, the models are predictive. That is, the results of the models are a prediction of both the number of crime trips originating in each zone and the number of crime trips ending in each zone (i.e., crimes occurring in a zone). Because the models are predictions, there is always error between the actual number and that predicted. As long as the error is not too large, the models can be useful for both analyzing the correlates of crime as well as being useful for forecasting or for simulating policy interventions.

Second, because the number of crimes attracted to the study jurisdiction will usually be greater than the number of crimes predicted for the origin zones, due primarily to crime trips coming from outside the origin areas, it is necessary to balance the productions and attractions. This is done in two steps. One, an estimate of trips coming from outside the study area (external trips) is added to the predicted origins as an 'external zone'. Two, a statistical adjustment is done in order to ensure that the total number of origins equals the total number of destinations. This is called *balancing* and is essential as an input into the second stage of crime travel demand modeling - trip distribution.

In the following discussion, first, the logic behind trip generation modeling is presented, including the calibration of a model, the addition of external trips in making a model, and the balancing of predicted origins and predicted destinations. Second, the mechanics of conducting the trip generation model within *CrimeStat* is discussed and illustrated with data from Baltimore County.

Modeling Trip Generation

The process of modeling trip generation is fairly well developed, at least with respect to ordinary trips. It proceeds through a series of logical steps that make up the aggregate trip generation model.

Trip Purpose

Trip generation modeling starts with the reasons behind travel. At an individual level, people make trips for a reason - to go to work, to go shopping, to go to a medical appointment, to go for recreation, or, in the case of offenders, to commit a crime. These are called *trip purposes*. Since there are a very large number of trip purposes, usually these are categorized into a few major groupings. In the case of the usual travel demand forecasting, the distinctions are *home-to/from-work* (or home-based work trips), *home-to/from-non-work* (or home-based non-work trips, e.g., shopping), and a *non-home trip* where neither the origin nor the destination are at the traveler's residence location (non-home-based trips).

Since the model has aggregated trips to a zone, the trip purposes are collections of trips from each origin zone to each destination zone. Thus, each zone produces a certain number of home-work trips, home-non-work trips, and non-home trips and each zone attracts a certain number of home-work trips, home-non-work trips, and non-home trips. This is the usual distinction that most transportation modeling organizations make. The trip purposes are documented during a large travel survey that asks individuals to fill out travel diaries for one or two days of travel. In the travel diaries, detailed information about each trip is documented - time of day, destination of trip, purpose of trip, travel modes used in making the trips, accompanying passengers, route taken, and time to complete the trip.

Aggregated Crime Trips

For crime trips, however, these distinctions are not very meaningful. There is very little information on how offenders make trips. One cannot just take a sample of offenders and ask them to complete a travel diary about how, when, and where the trip took place. With arrested offenders, it might be possible to produce such a diary, but both memory problems as well as legal concerns quickly make this an unreliable source of information. Therefore, as indicated in Chapter 26, a decision has been made to reference all trips with respect to the residential home location. All crime trips are analyzed as *home-crime* trips.

However, other distinctions can be made. The most obvious is by type of crime. There are robbery trips, burglary trips, vehicle theft trips, and so forth. Similarly, distinctions can be made by travel time such as afternoon trips or evening trips. However, the sample size will

decrease with greater distinctions. Logically, one can divide a sample into a very large number of important distinctions (e.g., afternoon burglary trips involving two or more offenders). However, this reduces the sample size and increases the error in estimation, particularly at the trip distribution and subsequent stages.

An important point that distinguishes the aggregate demand types of travel demand models, as is being implemented here, and the newer generation of activity-based trips is that there are no *linked trips* with the aggregate approach (Pribyl & Goulias, 2005). If an offender first steals a car, then uses the car to rob a grocery store followed by a burglary, the aggregate approach models this as three separate trips, rather than as a series of three linked crime trips (which the activity-based models do). This is a deficiency with the aggregate travel demand model. In order to make the aggregate models work, each trip is considered independent of any other trip. While this is not realistic behaviorally, since we know that many crimes are committed in sequence as part of a single journey (or tour), the zonal approach does limit the underlying logic of crime trips. Nevertheless, the aggregate approach can be very useful as long as it implemented consistently. With the current state of activity-based modeling, there is not yet any evidence that they produce more accurate predictions than the cruder, aggregate approach (Culp & Lee, 2005).

Correlates of Crime

Any trip has contextual correlates associated with it. It is well documented that the likelihood of making a trip (crime or otherwise) is not equal across areas of a metropolitan region. There are age and gender correlates of travel, socioeconomic correlates of travel, and land use correlates of travel; the latter are usually associated with trip purposes (e.g., retail areas attract shopping trips).

The trip generation model being implemented in this version of *CrimeStat* is an aggregate model. Thus, the predictors are aggregate, rather than behavioral, in nature, as discussed in Chapter 25. They are correlates of trips, not necessarily the *reasons* for the trips. For example, typically population is the best predictor of trips. Zones with many persons will produce, on average, more crime trips than zones with fewer persons. The observation is not a reason, but is simply a by-product of the size of the zone. Similarly, low-income zones will tend to produce, on average, more crime trips than wealthier zones; again, this is not a reason, but a correlate of the characteristics that might contribute to individual likelihoods for committing crimes.

As mentioned in Chapter 25, there are a number of different variables that could be used for prediction, although population (or a proxy for population, such as households), income or poverty, and land use variables would be the most common (NCHRP, 1998).

Theoretical Relevance of the Variables

In general, the variables that are selected should be empirically stable and theoretically meaningful. That is, they should be stable variables that do not change dramatically from year to year. They should be reliably measured so that an analyst can depend on their values. Finally, they should be meaningful in some ways. That is, they should be plausible enough that both crime analysts and researchers and informed outsiders should agree that the relationship is plausible. The variables either should have been demonstrated to be predictors in earlier research or else to be so correlated with known factors as to be considered meaningful proxies.

Spurious correlates

On the other hand, if a variable is either a correlate of a known predictor or idiosyncratic, then it is liable not be believed. For example, the number of taxis usually correlates with the amount of employment since taxis tend to ply commercial areas for their trade. Adding the number of taxis in a predictive model is liable to produce significant statistical effects in predicting crime destinations. However, few persons are going to believe that this is a real factor since it is understood to be a correlate of a more structural variable.

Idiosyncratic variables are those that appear in unique situations. For example, in some cities, adjacency to a freeway is a correlate of crime origins (e.g., in Baltimore County where low income populations live) whereas in other cities, it is a correlate of crime destinations (e.g., in Houston where there are frontage roads with major commercial strips that attract crimes). The variables may be real predictors. However, the analyst or researcher will have difficulty persuading others to believe in the model, at least until the results can be replicated.

In other words, what is required for the model is a set of reasonable correlates of crime trips that would be plausible and stable over time. It is an ecological model, not a behavioral one.

Social Disorganization Variables

There is a very large literature on the predictors of crime, typically following from the social disorganization literature (for example, Park &Burgess, 1924; Thrasher, 1927; Shaw & McKay, 1942; Newman, 1972; Ehrlich, 1975; Cohen & Felson, 1979; Wilson & Kelling, 1982; Stack, 1984; Messner, 1986; Chiricos, 1987; Kohfeld & Sprague, 1988; Bursik & Grasmick, 1993; Hagan & Peterson, 1994; Fowles & Merva, 1996; Bowers & Hirschfield, 1999 among many other studies). Much of this literature identifies correlates that are associated with crime incidents. Among the factors that have been associated with crime and delinquency at an aggregate geographical level are poverty, low income households, overcrowding, substandard

housing, low education levels, single-parent households, high unemployment, minority and immigrant populations.[1]

Statistical Problems with Predictor Variables

Multicollinearity among the independent variables

There are two statistical problems associated with using these variables as predictors. The first is the high degree of overlap between the variables. Zones that have high poverty levels typically also have low household income levels, higher population densities, substandard housing, a high percentage of renters, and higher proportion of minority and immigrant populations. In a regression model, this overlap causes a condition known as *multicollinearity*. Essentially, the independent variables correlate so highly among themselves that they produce ambiguous, and sometimes strange, results in a regression model. For example, if two independent variables are highly correlated, frequently one will have a positive coefficient with the dependent variable while the other will have a negative coefficient; conversely, they sometimes can cancel each other out. Chapter 17 discussed multicollinearity and provided an example that showed correlated independent variables can cancel each other out. Thus, in spite of the correlates with crime levels, in a model it is usually best to eliminate *co-linear* variables. The result is that simple variables usually end up being the most straightforward to use (population, median household income) with many of the subtle, but theoretically relevant, variables typically dropping out of the equation.

Failure to distinguish origins from destinations

Second, in much of this literature, however, there is not a clear distinction between origin predictors and destination predictors. That is, in most cases, the correlates of crimes were identified but it is often unclear whether these correlates are associated with the neighborhoods of the offenders (origins) or the locations where the crimes occur (destinations). This can result in a set of vague correlates without clear direction about whether the variables are associated with producing or attracting conditions. In fact, in much of the early literature on social disorganization, it was implicitly assumed that crimes are produced in the neighborhoods where the offenders lived, a linkage that is increasingly becoming disconnected. For modeling crime trips, however, it is essential that the predictors of origins be kept separate from the predictors of destinations.

---

[1] Note that a correlation at an aggregate level does not necessarily imply a correlation at the individual level. As has been noted frequently, the vast majority of people do not commit serious crimes and that most crimes are committed by a small proportion of the population (Ratcliffe, 2008).

Accuracy and Reliability

A trip generation model should be accurate and reliable. *Accuracy* means that the model should replicate as closely as possible the actual number of trips originating or ending in zones and that there should be no bias (which is a systematic under- or over-estimating of trips). *Reliability* means that the amount of error is minimized.

These criteria have two implications which are somewhat at odds. First, we have to choose models that replicate as closely as possible the number of trips originating or ending in a zone. In general, this would be a model that had the highest overall predictability. But, second, we have to choose models that minimize total prediction errors. This allows a model to replicate the number of trips for as many zones as possible. The two criteria are somewhat contradictory because crime trips are highly skewed. That is, a handful of zones will have a lot of crimes originating or ending in them while most zones will have few or no crimes. The zones with the most crimes will have a disproportionate impact on the final model. Thus, a model that obtains as high a prediction as possible (i.e., highest log-likelihood or R^2) may actually only predict accurately for a few zones and may be very wrong for the majority.

The strategy, therefore, is to obtain a model that balances high predictability but by keeping the total prediction error low.

Count Model

Another element of the model is that the trip generation model is for *counts* (or volumes), not for rates. The model predicts the number of crimes originating in each origin zone and the number of crimes occurring in each destination zone. The model could be constructed to predict rates, but normally it is not done. For most travel demand modeling, as mentioned in Chapter 25, the model predicts the *number* of trips originating or ending in a zone. Thus, there is a *crime production* model that predicts the number of crimes originating in each zone and a *crime attraction* model that predicts the number crimes

Approaches Toward Trip Generation Modeling

Trip Tables

There are two classic approaches to trip generation modeling. The first uses a *trip table* (sometimes called a cross-classification table or a category analysis). A trip table is a cross-classification matrix. Several predictive variables are divided into categories (e.g., three level of household income; four levels of vehicle ownership; three levels of population density) and a mean number of trips is estimated for each cell, usually from a survey. For example, a survey of

household income might show the relationship between household income and the number of trips taken by individuals of the households. Based on a sample, estimates of the *average number of trips per person* can be obtained for each income level (e.g., 3.4 trips per day for persons from low income households; 4.5 trips per day for persons from median income households; 6.7 trips per day for persons from high income households). These variables are further subdivided into two-way or three-way cross-tabulation tables (e.g., low income and medium vehicle ownership; low income and high vehicle ownership). Table 27.1 illustrates a *possible* trip table model involving two variables. In practice, three or four variables are used.

The main reason that trip tables are used in a trip generation model is because of the non-linear nature of trips. Predictive variables are usually not linear in their effects on the number of trips. Thus, unless a sophisticated non-linear model is used, sizeable error can be introduced in a prediction. It is usually safer to use a trip table approach (Ortuzar & Willumsen, 2001). There are some major handbooks on the topic (Henscher & Button, 2002; ITE, 2003). In fact, the Institute of Transportation Engineers publishes a large handbook that gives extensive trip production and trip attraction tables by detailed land uses (ITE, 2003). These tables are often used in formal environmental review processes for site analysis and are frequently accepted by courts in litigation. They are not without their problems, however, and there have been numerous critiques of the tables (Shoup, 2002; NCHRP, 1998). They also cannot be used in a travel demand model and will produce erroneous results.

The problem for crime analysis, however, is that it is impossible to obtain these data. One cannot ask a sample of offenders how many crimes they undertake each day in order to estimate the mean expectations for a table. Thus, one has to adopt a more indirect approach in modeling crime productions and attractions.

A second problem with the trip table approach is its use of zonal data. While it could be applied to zonal data (e.g., using median household income and average vehicle ownership in Table 27.1 instead of individual household income and vehicle ownership), this type of approach is prone to ecological inference errors and could be very wrong (Freedman, 1999; Langbein & Lichtman, 1978). There is no guarantee that the splitting of two aggregate variables (essentially, the cross-product of their marginal probabilities) will produce an accurate trip estimate; often, such an approach leads to very wrong results.

Further, such an approach requires interpretation and some degree of arbitrariness. For example, how does one subdivide median household income? One person might interpret it slightly differently than another; unlike simple numerical counts (e.g., 0 vehicle ownership;

Table 27.1:
Illustration of Possible Trip Table Approach to Trip Generation
Average Trips per Adult, Age 16+

		Household income		
		Low	Medium	High
Vehicle Ownership	0-1	3.2	4.6	6.7
	2+	5.4	7.8	8.1

1 vehicle ownership; 2 vehicle ownership), there is too much variability in categorizing variables at the zonal level.[2]

Linear/OLS Regression Modeling

The second approach is to use a *regression* framework. In this approach, the number of crimes either originating or ending in each zone is estimated from zone characteristics using a regression model. This can be written in a generalized linear model ('link' function) form (see Chapter 16):

$$f(Y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots . \beta_K X_K + \epsilon \qquad (27.1)$$

This equation says that some function of the mean number of crimes, $f(Y_i)$, either originating or ending in zone I, is a linear function of a number of independent variables, X_1, X_2, X_3,....X_K for these zones; there are *K* independent variables plus a possible constant. There is also an error term which represents the discrepancy between the actual observation and what the model predicts. This is sometimes called *residual error* since it is the difference between the observed and predicted values (O_i - Y_i). The function is unspecified and can be non-linear.[3]

The traditional approach to regression modeling assumed that the independent variables are linear in their effect on the dependent variable. Thus,

---

[2] There is also subjectivity in subdividing variables at an individual level. For example, household income levels can be subdivided in different ways. However, with aggregate data, all variables have to be subdivided arbitrarily whereas with individual level data, typically only income is done this way.

[3] Some statisticians often refer to the number of *parameters* that have to be estimated in an equation, not just the number of independent variables. In most regression models, for example, there are K+1 parameters that are estimated - coefficients for the K independent variables and a constant term. In this text, K refers to the number of independent variables, not estimated parameters.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots . \beta_K X_K + \epsilon \quad (27.2)$$

In this model, there are K independent variables and one constant term (β_0, sometimes called α) that needs to be estimated. For each zone, *i*, each of the independent variables has a weight associated with it (the coefficient, β). The product of the value of the independent variable times its weight represents its *effect*. The individual effects of each of the K independent variables are summed to produce an overall estimate of the dependent variable, Y.

The method for estimating this equation usually minimizes the sum of the squares of the residual errors. Hence, the procedure is called *Ordinary Least Squares* (or OLS). If the equation is correctly specified (i.e., the dependent variable is normally distributed and all relevant variables have been included), the error term, ε, will be normally distributed with a mean of 0 and a constant variance, σ^2.

Problems with OLS Regression Modeling

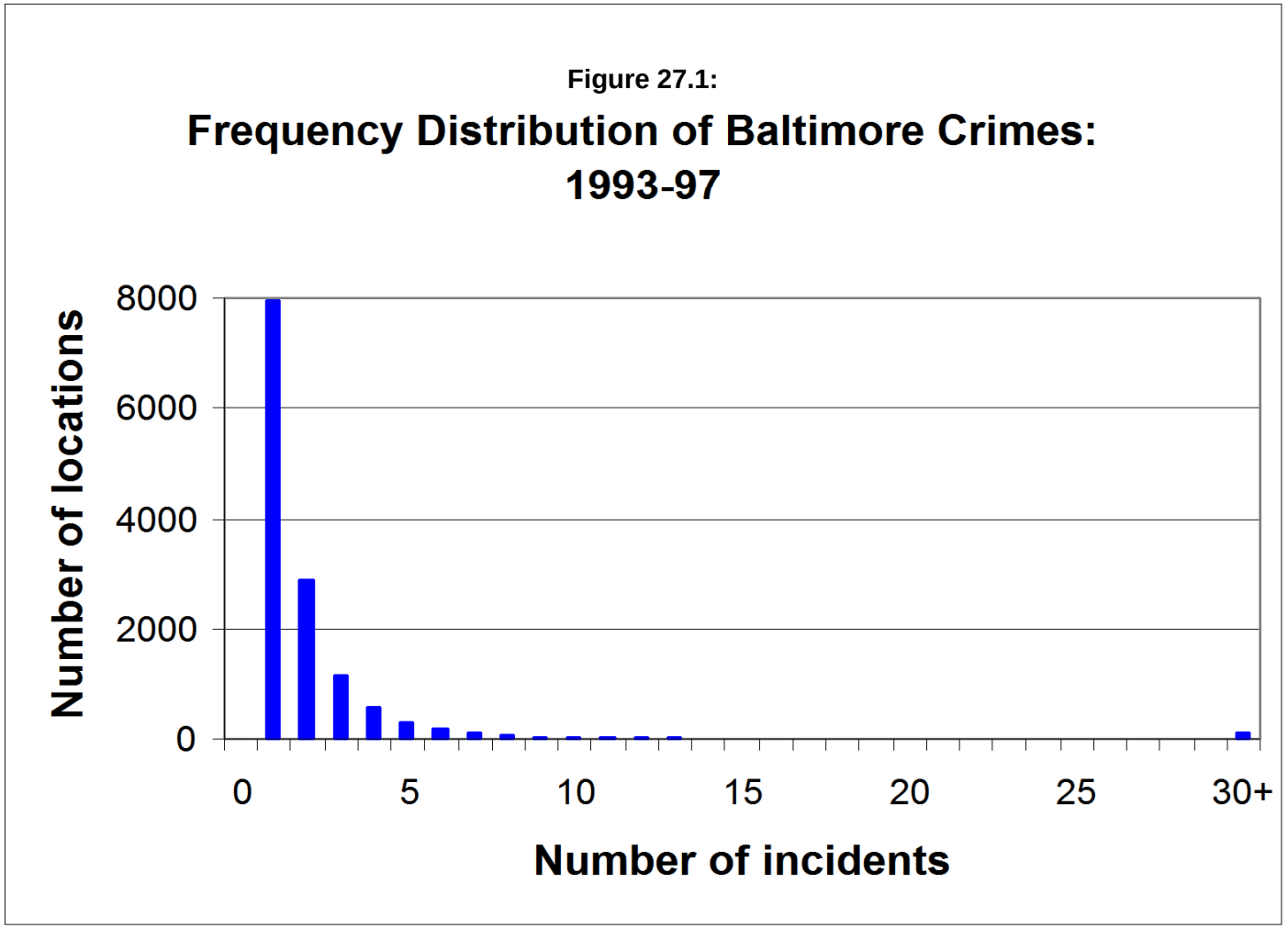
However, there are a number of major problems associated with OLS regression modeling. These were discussed in Chapter 15 (15.16-15.19). To repeat, there are five major problems with the OLS model.

Skewness of crime events

First, crime events are extremely statistically skewed. Some locations have a much higher likelihood of a crime event (either an origin or a destination) than others. Figure 27.1 below shows the number of crimes from 1993 to 1997 in Baltimore County that occurred at each location. That is, the graph shows the number of incidents that occurred at every location, plotted in decreasing order of frequency. Thus, there were 7,965 locations where only one crime occurred between 1993 and 1997. There were 2,878 locations where two crimes occurred in that period. There were 1,138 locations where three crimes occurred in that period. At the other end of the spectrum, there were 332 locations that had 10 or more crimes during the period and there were 97 locations that had 30 or more crimes occur. If we add to this the very large number of locations where no crimes occurred, the unequal likelihoods of crime by location is even more dramatic. In other words, the data are highly skewed with respect to the frequency of crimes. Most locations either had no crimes occur or very few, while a few locations had many crimes occur.

Aggregating crimes into zones tends to reduce *some* of the skewness. For example, grouping the crimes by origin traffic analysis zone (TAZ) reduced it a little bit. Nineteen of the 525 origin zones in Baltimore County and Baltimore City did not have any crimes occur in them while 15 zones had only one crime occur. Six zones had two crimes originate from them while 8

Figure 27.1:


Frequency Distribution of Baltimore Crimes: 1993-97
Number of locations
8000
6000
4000
2000
0
0
5
10
15
20
25
30+
Number of incidents

zones had three crimes originate from them. At the other end, 1 zone had 738 crimes originate from it and another zone had 533 originate from it. Of the 525 origin zones, 155 had 100 or more crime events. Similar results are found for the destination zones. Figure 27.2 graphs the distribution of origins and destinations by TAZ's in bins of 50 incidents each.

Skewness in the dependent variable usually makes the final model biased and unreliable. Particularly if the skewness is positive (i.e., a handful of cases have very large values), the resulting regression coefficients will reflect the cases with the highest values rather than represent all the cases with approximately equal weights. These so-called 'outliers' can overwhelm a regression equation. In an extreme case, a very large outlier may totally determine the model.[4]

Skewness makes prediction difficult. The OLS model assumes that each independent variable contributes to the dependent variable at an arithmetic rate; there is a constant slope such that a one unit change in the independent variable is associated with a constant change in the dependent variable. With skewness, on the other hand, such a relationship will not be found. Large changes in the independent variable will be necessary to produce small changes in the dependent variable, but the effect is not constant. In other words, the OLS model typically cannot explain the non-linear changes in the dependent variable.[5]

Negative predictions

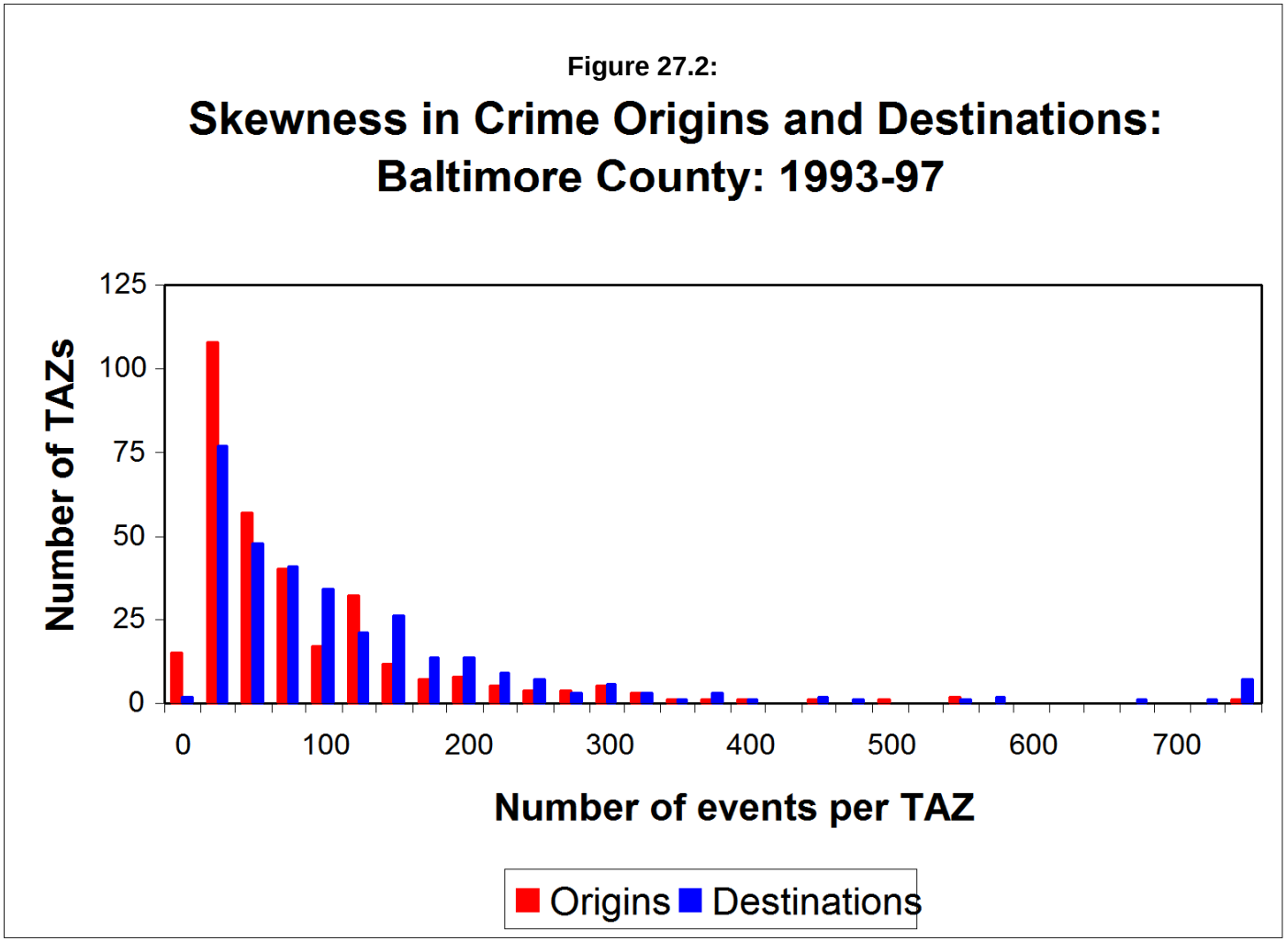
A second problem with OLS is that it can have negative predictions. With a count variable, such as the number of crimes originating or ending in a zone, the minimum number is zero. That is, the count variable is always *positive*, being bounded by 0 on the lower limit and some large number on the upper limit. The OLS model, on the other hand, can produce negative predicted values since it is additive in the independent variables. This clearly is illogical and is a major problem with data that are skewed. If the most common value is close to zero, it is very possible for an OLS model to predict a negative count.

---

4 For example, an experiment with 100 cases was created with a progressing dependent variable and a **random** independent variable (i.e., the independent variable had its value selected randomly). The dependent variable progressed from 1 to 100. For the first 99 cases, the independent variable took values from 0.12 to 9.9, randomly assigned. The correlation between these two variables for the first 49 cases was 0.04. However, for the 100th case, the independent variable was given a value of 100. The correlation between the two variables now shot up to 0.17. Even though the F-test for this was not significant, it represented a sizeable jump. Replacing one other independent value with a 50 caused the correlation to jump to 0.23, which was statistically significant. In other words, two outliers caused a random series to appear significant!

5 It is possible to transform the independent variable into a non-linear predictor, for example by taking the log of the independent variable or raising it to some power (e.g., X^2). However, this will not solve the other problems associated with OLS, namely negative and non-summative predictions.

Figure 27.2:


Skewness in Crime Origins and Destinations:
Baltimore County: 1993-97
Number of TAZs
125
100
75
50
25
0
0
100
200
300
400
500
600
700
Number of events per TAZ
Origins
Destinations

Non-consistent summation

A third problem with OLS models is that the sum of the input data values do not necessarily equal the sum of the predicted values. Since the estimate of the constant and coefficients is obtained by minimizing the sum of the squared residual errors, there is no balancing mechanism to require that they add up to the same as the input values. For a trip generation model in which the number of predicted origins must equal the number of predicted destinations (after adding in the number of predicted external trips), this can be a big problem. In calibrating the model, adjustments can be made to the constant term to force the sum of the predicted values to be equal to the sum of the input values. But in applying that constant and coefficients to another data set, there is no guarantee that the consistency of summation will hold. In other words, the OLS method cannot guarantee a consistent set of predicted values.

Non-linear effects

A fourth problem with the OLS model is that it assumes the independent variables are linear in their effect. If the dependent variable was normal or relatively balanced, then a linear model might be appropriate. But, when the dependent variable is highly skewed, as is seen with these data, typically the additive effects of each component cannot usually account for the non-linearity. Independent variables have to be transformed to account for the non-linearity and the result is often a complex equation with non-intuitive relationships.[6] It is far better to use a non-linear model for a highly skewed dependent variable.

Uneven residual errors

The final problem with an OLS model and a skewed dependent variable is that the model tends to over- or under-predict the correct values, but rarely comes up with the correct estimate. With skewed data, typically an OLS equation produces non-constant residual errors. That is, one of the major assumptions of the OLS model is that all relevant variables have been included. If that is the case, then the errors in prediction (the residual errors - the difference between the observed and predicted values) should be uncorrelated with the predicted value of the dependent variable. Violation of this condition is called *heteroscedasticity* because it indicates that the residual variance is not constant. The most common type is an increase in the residual errors with

---

[6] For example, to account for a skewed dependent variable, one or more of the independent variables have to be transformed with a non-linear operator (e.g., log or exponential term). When more than one independent variable is non-linear in an equation, the model is no longer easily understood. It may end up making reasonable predictions for the dependent variable, but it is not intuitive and not easily explained to non-specialists. It is possible to transform the independent variable into a non-linear predictor, for example by taking the log of the independent variable or raising it to some power (e.g., X^2). However, this will not solve the other problems associated with OLS, namely negative and non-summative predictions.

higher values of the predicted dependent variable. That is, the residual errors are greater at the higher values of the predicted dependent variable than at lower values (Draper & Smith, 1981, 147).

A highly skewed distribution tends to encourage this. Because the least squares procedure minimizes the sum of the squared residuals, the regression line balances the lower residuals with the higher residuals. The result is a regression line that neither fits the low values or the high values. For example, motor vehicle crashes tend to concentrate at a few locations (crash hot spots). In estimating the relationship between traffic volume and crashes, the hot spots tend to unduly influence the regression line. The result is a line that neither fits the number of expected crashes at most locations (which is low) nor the number of expected crashes at the hot spot locations (which are high). The line ends up over-estimating the number of crashes for most locations and under-estimating the number of crashes at the hot spot locations.

Poisson Regression Modeling

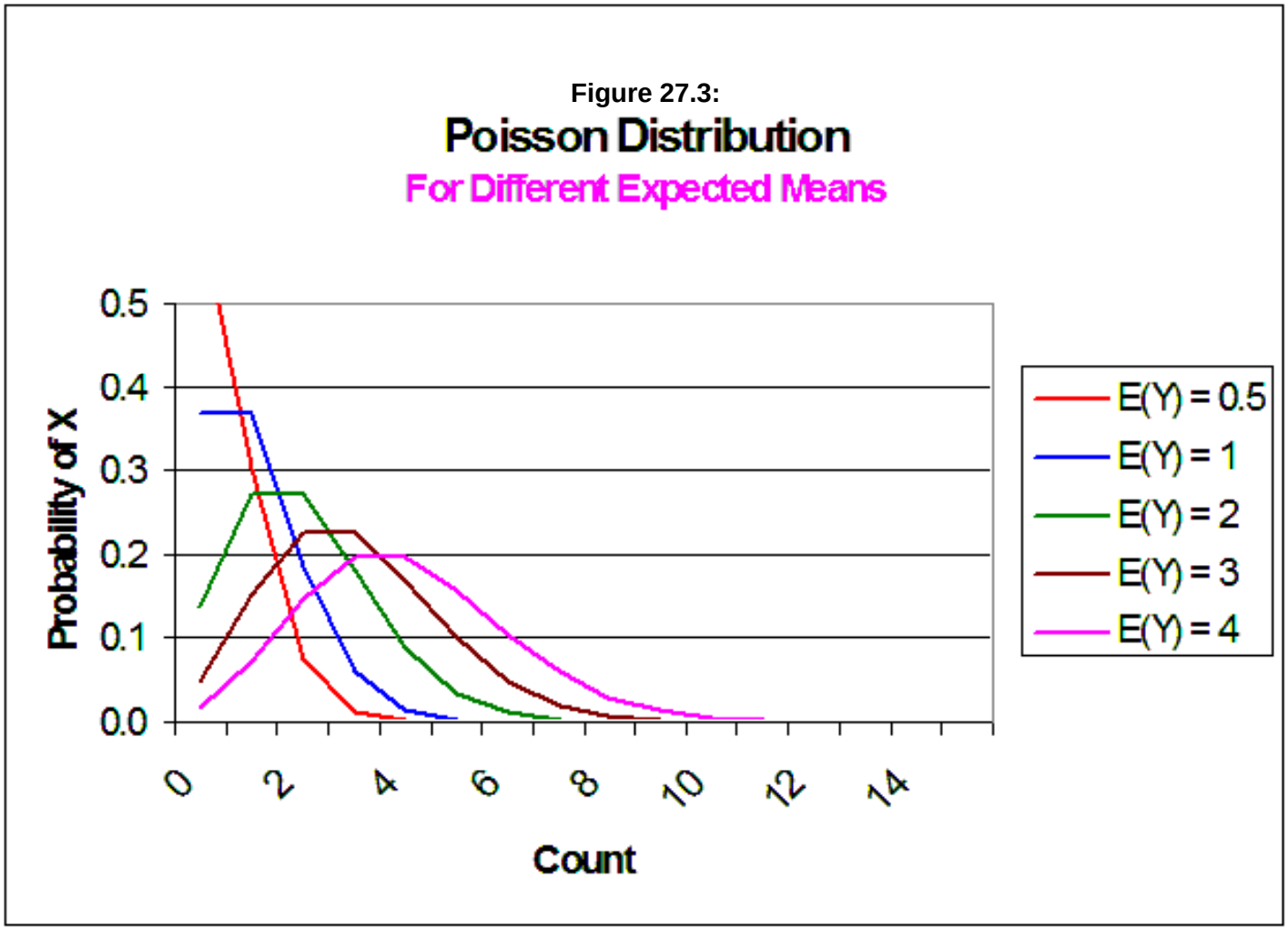
Poisson regression is a non-linear modeling method that overcomes some of the problems of OLS regression. It is particularly suited to count data (Cameron & Trivedi, 1998). In the model, the number of events is modeled as a Poisson random variable:

$$E(Y_i) = \frac{e^{-\lambda}\lambda^{Y_i}}{Y_i!} \qquad (27.3)$$

where Y_i is the count for one group or class, i, λ is the mean count over all groups, and **e** is the base of the natural logarithm. The distribution has a single parameter, λ, which is both the mean and the variance of the function.

The 'law of rare events' assumes that the total number of events will approximate a Poisson distribution *if* an event occurs in any of a large number of trials but the probability of occurrence in any given trial is small (Cameron & Trivedi, 1998). Thus, the Poisson distribution is very appropriate for the analysis of rare events such as crime incidents (or motor vehicle crashes or rare diseases or any other rare event). The Poisson model is not particularly good if the probability of an event is more balanced; for that, the normal distribution is a better model as the sampling distribution will approximate normality with increasing sample size. Figure 27.3 illustrates the Poisson distribution for different expected means.

Figure 27.3:


Poisson Distribution
For Different Expected Means
Probability of X
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
14
Count
E(Y) = 0.5
E(Y) = 1
E(Y) = 2
E(Y) = 3
E(Y) = 4

The mean can, in turn, be modeled as a function of some other variables (the independent variables). Given a set of observations on dependent variables, X_{ki} (X_1, X_2, X_3,...,X_K), the *conditional mean* of Y_i can be specified as an exponential function of the X's:

$$E(y_i \mid \mathbf{x}_i) = \lambda_i = e^{\mathbf{x}_i^T \boldsymbol{\beta}} \tag{27.4}$$

where X_{ki} is a set of independent variables, β is a set of coefficients, and **e** is the base of the natural logarithm.. Now, the conditional mean (the mean controlling for the effects of the independent variables) is non-linear. Equation 27.4 is sometimes written as:

$$Ln(\lambda_i) = X_{ki}\beta \tag{27.5}$$

and is known as the *loglinear* model. In more familiar notation, this is written as:

$$Ln\big((\lambda_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_K X_K = \beta_0 + \textstyle\sum_{K=1}^{K} (\beta_K X_K\big) \tag{27.6}$$

That is, the natural log of the mean is a function of K random variables and a constant.

Note, that in this formulation, there is not a random error term. The data are assumed to reflect the Poisson model. There can be residual errors, but these are assumed to reflect an incomplete specification (i.e., not including all the relevant variables. Also, since the variance equals the mean, it is expected that the residual errors should increase with the conditional mean. That is, there is inherent heteroscedasticity (Cameron & Trivedi, 1998). This is very different than an OLS where the residual errors are expected to be constant.

The model is estimated using a maximum likelihood procedure, typically the Newton-Raphson method. In Appendix B, Luc Anselin presents a more formal treatment of both the OLS and Poisson regression models

Advantages of the Poisson Regression Model

The Poisson model overcomes some of the problems of the OLS model. First, the Poisson model has a minimum value of 0. It will not predict negative values. This makes it ideal for a distribution in which the mean or the most typical value is close to 0. Second, the Poisson is a fundamentally skewed model; that is, it is non-linear with a long 'right tail'. Again, this model is appropriate for counts of rare events, such as crime incidents.

Third, because the Poisson model is estimated by either maximum likelihood or Markov Chain Monte Carlo (MCMC; see chapters 16 and 17), the estimates are adapted to the actual data.

In practice, this means that the sum of the predicted values is virtually identical to the sum of the input values, with the exception of very slight rounding off error. In the subsequent balancing of the predicted origins and the predicted destinations, this leads to a more stable estimate since the only difference between the predicted origins and predicted destinations is the number of trips that come from outside the study area (external trips). Since the external trips are added to the predicted origins, the balancing operation is less prone to adjustment error.

Fourth, compared to the OLS model, the Poisson model generally gives a better estimate of the number of crimes for each zone. The problem of over- or under-estimating the number of incidents for most zones with the OLS model is usually lessened with the Poisson. When the residual errors are calculated, generally the Poisson has a lower total error than the OLS.

In short, the Poisson model has some desirable statistical properties that make it very useful for predicting crime incidents (origins or destinations).

Problems with the Poisson Regression Model

On the other hand, the Poisson model is not perfect. The primary problem is that count data are usually *over-dispersed* but occasionally can be *under-dispersed*.

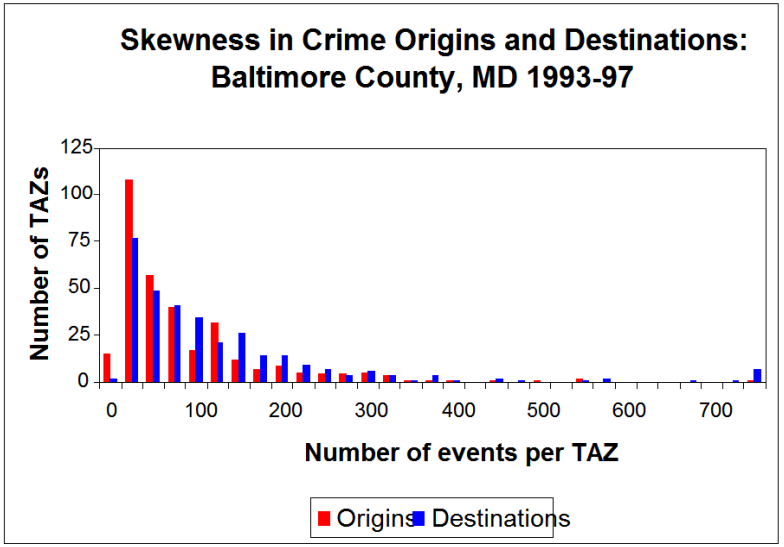
Over-dispersion in residual errors

In the Poisson distribution, the mean equals the variance. In a Poisson regression model, the mathematical function, therefore, equates the conditional mean (the mean controlling for all the predictor variables) with the conditional variance. However, most real data are over-dispersed; the variance is generally greater than the mean. Figure 27.4 shows the distribution of Baltimore County and Baltimore City crime origins and Baltimore County crime destinations by TAZ (repeat of Figure 27.2) and also indicates the variance-to-mean ratio of each variable. For the origin distribution, the ratio of the variance to the mean is 14.7; that is, the variance is 14.7 times that of the mean! For the destination distribution, the ratio is 401.5!

In other words, the variance is many times greater than the mean. Most real-world count data are similar to this; the variance will usually be a lot greater than the mean. What this means in practice is that the residual errors - the difference between the observed and predicted values for each zone, will be greater than what is expected. The Poisson model calculates a standard error as if the variance equals the mean. Thus, the standard error will be underestimated using a Poisson model and, therefore, the significance tests (the coefficient divided by the standard error) will be greater than it really should be. This would have the effect of identifying variables as being more statistically significant in a model than what they actually should be. In other words, in a Poisson

Figure 27.4:

Over-dispersion



Origins:

Mean = 75.8

Variance = 7848.8

Ratio of variance to mean = 14.7

Destinations:

Mean = 129.1

Variance = 51,849.1

Ratio of variance to mean = 401.5

regression model, we would end up selecting variables that really should not be selected because we think they are statistically significant when, in fact, they are not.

Another problem with the Poisson, which is true for most of the common regression methods, is the lack of a spatial predictor component. For these, the MCMC Poisson models, discussed in Chapter 17 can be used. Also, in the crime travel demand model, spatial interaction is explicitly incorporated during the second stage of the model - trip distribution. Thus, any errors introduced in the first stage - trip generation, are usually compensated for during the second. Nevertheless, the inclusion of a spatial component in a regression model will generally improve the prediction.

Dispersion correction parameter

There are a number of methods for correcting the over-dispersion in a count model. Most of them involve modifying the assumption of the conditional variance equal to the conditional mean. For example, the negative binomial model assumes a Poisson mean but a gamma-distributed variance term (Cameron & Trivedi, 1998, 62-63; Venables & Ripley, 1997, 242-245). That is, there is an unobserved variable that affects the distribution of the count. There are several interpretations of the negative binomial (see Boswell & Patil, 1970) but the most common is to assume that there are mixtures of distinct Poisson distributions that make up the real distribution.

The negative binomial model has a Poisson mean but with a 'longer tail' variance function and is usually preferred for over-dispersed data sets, such as typical with crime data. In Appendix C, Dominique Lord and Byung-Jung Park present a formal treatment of the negative binomial model. Other adjustments that can be made include the Poisson-Lognormal model (which can be estimated in *CrimeStat*; see Chapter 17) and the zero-inflated Poisson model assumes a Poisson function combined with a degenerate function with a probability of 1 for zero counts (Hall, 2000). These generally produce better estimates that the simple Poisson especially if a spatial component is added.

There is a simple linear correction for over-dispersion that frequently works, called the ***NB1*** model (Cameron & Trivedi, 1998, 63-65). The model proceeds in two steps. In the first, the Poisson model is fitted to the data and the degree of over- (or under-) dispersion is estimated. The dispersion parameter is defined as:

$$\Phi = \frac{1}{N-K-1}\sum_{i=1}^{N}\frac{(Y_i-P_i)^2}{P_i} \qquad (27.7)$$

where N is the sample size, K is the number of independent variables, Y_i is the observed number of events that occur in zone i, and P_i is the predicted number of events for zone i. The test is similar

to an average Chi-square in that it takes the square of the residuals (Y_i - P_i) and divides it by the predicted values, and then averages it by the degrees of freedom. The dispersion parameter is a standardized number. A value greater than 1.0 indicates over-dispersion while a value of less than 1 indicates under-dispersion (which is rare, though possible). A value of 0 indicates *equidispersion* (or the variance equals the mean).

In the second step, the Poisson standard error is multiplied by the square root of the dispersion parameter to produce an *adjusted standard error*:

$$SE_{adj} = SE * \sqrt{\Phi} \quad (27.8)$$

The new standard error is then used in the t-test to produce an adjusted t-value. Cameron and Trivedi (1998) have shown that this adjustment produces results that are almost identical to that of the negative binomial, but involving fewer assumptions. Chapter 16 discussed the NB1 model in more depth.

The point is that the Poisson model needs to be adjusted for over-dispersion. CrimeStat provides a number of regression tools for accounting over-dispersion and which can also include a spatial autocorrelation adjustment. Chapters 16 and 17 provide information on these models.

Under-dispersion in residual errors

Occasionally, a data set will be under-dispersed, meaning that the conditional variance is substantially lower than the mean. As a rough approximation, Cameron and Trivedi (1998) suggest that if the raw variance-to-mean ratio is less than 2.0, then most likely the model will show under-dispersion with the conditional mean. If the under-dispersion is slight, then the NB1 model can be used to adjust the standard errors. If it is substantial, however, then other models have to be considered. See Chapter 17 for more details.

Diagnostic Tests

There are a number of diagnostics tests that are used in a regression framework.

Skewness Tests

First, there are tests of skewness in the dependent variable. As mentioned above, the OLS model cannot be applied to data that are highly skewed. If they are skewed, a non-linear model, such as the Poisson, must be used. Therefore, it is essential to evaluate the degree of skewness.

A commonly used measure of skewness is the g statistic (Microsoft, 2012):

$$Skewness\ (g) = \frac{N}{(N-1)(N-2)} \sum_{i=1}^{N} [\frac{(X_i - \overline{X})}{s}]^3 \quad (27.9)$$

where N is the sample size, X_i is observation i, $\bar{X}$ is the mean of X, and s is the sample standard deviation (corrected for degrees of freedom):

$$s = \sqrt{\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N-1}} \quad (27.10)$$

The standard error of skewness (SES) can be approximated by (Tabachnick & Fidell, 1996):

$$SES = \sqrt{\frac{6}{N}} \quad (27.11)$$

An approximate Z-test can be obtained from:

$$Z(g) = \frac{g}{SES} \quad (27.12)$$

Thus, if Z is greater than +1.96 or smaller than -1.96, then the skewness is significant at the $p \leq .05$ level.

As an example, for the data on the origins of crimes by TAZ in Baltimore County, we have:

$$\bar{X} = 75.108 \quad (27.13)$$

$$s = 96.017 \quad (27.14)$$

$$N = 325 \quad (27.15)$$

$$\sum_{i=1}^{N} [\frac{(X_i - \overline{X})}{s}]^3 = 898.31 \quad (27.16)$$

Therefore,

$$g = \frac{325}{324*323} * 898.391 = 2.79 \quad (27.17)$$

$$SES = \sqrt{\frac{6}{325}} = 0.136 \quad (27.18)$$

$$Z(g)=20.51 \quad (27.19)$$

The Z of the g value shows the data are highly skewed as was, of course, already known.

Likelihood Ratio Test

Second, there are tests of the overall model. In a maximum likelihood framework, the first test is of the *log-likelihood* function. A *likelihood* function is the joint density of all the observations, given a value for the parameters, β, and the variance, σ^2. The log-likelihood is the natural log of this product, or the sum of the logs of the individual densities. For the OLS model, the log-likelihood is:

$$L = 1\left(\frac{N}{2}\right)\ln(2\pi) - \left(\frac{N}{2}\right)\ln(\sigma^2) - \left(\frac{\sigma}{2}\right) - 0.5\frac{(Y_i - X_{ki}\beta_k)^2}{\sigma^2} \quad (27.20)$$

where N is the sample size, σ^2 is the variance, Y_i is the observed number of events for zone i, and $X_{ki}\beta_k$ is a series of K independent predictors multiplied by their coefficients.

In the Poisson model, the log-likelihood is:

$$L = \sum_{i=1}^{N}[-\lambda_i + Y_i X_{ki}\beta_k - lnY_i!] \quad (27.21)$$

where λ_i is the conditional mean for zone i, Y_i is the observed number of events for zone i, and Y_i $X_{ki}\beta_k$ is a cross-product of the observed events times the K independent predictors multiplied by their coefficients. As mentioned above, Luc Anselin provides a more detailed discussion of these functions in Appendix B.

Since the maximum likelihood method achieves the model with the highest log-likelihood, the log-likelihood is a negative number. Even though the model with the highest log-likelihood is considered 'best', it is not an intuitive number. Consequently, the *Likelihood Ratio* compares the log-likelihood of the regression model with the log-likelihood that would be obtained if only the mean number of counts was taken. This latter log-likelihood is:

$$L_R = -N\bar{Y} + [\ln(\bar{Y})\sum_{i=1}^{N} Y_i] - \sum_{i=1}^{N} Y_i! \quad (27.22)$$

The Likelihood Ratio test is:

$$LR = 2(L - L_R) \tag{27.23}$$

where L is the model log-likelihood and L_R is the log-likelihood of the mean count. The Likelihood Ratio is twice the difference between log-likelihood values of the regression and mean models respectively. It follows a χ^2 distribution with K degrees of freedom (where K is the number of independent variables).[7]

Adjusted likelihood ratio

The Likelihood Ratio is a more intuitive index since it is a Chi-square test. However, it is prone to the problem of all regression methods of over-fitting - the more independent variables are added to the model, the higher is the Likelihood Ratio. Consequently, there are several methods that adjust for the number of parameters fit. One is the Akaike Information Criterion (AIC) which is defined as:

$$AIC = -2L + 2(K + 1) \tag{27.24}$$

where L is the log-likelihood and K is the number of independent variables. A second one is the Bayesian Information Criterion/Schwartz Criterion (BIC/SC), which is defined as:

$$\text{BIC/SC} = \text{-2L+[(K+1)ln(N)]} \tag{27.25}$$

These two measures penalize the number of parameters added in the model, and reverse the sign of the log-likelihood (L) so that the statistics are more intuitive. The model with the lowest BIC/SC value is 'best'.

R-square Test

The most familiar test of an overall model is the R-square (or R^2) test. This is the percent of the total variance of the dependent variable accounted for by the model. More formally, it is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - P_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{27.26}$$

[7] Note, in Appendix B Luc Anselin uses K for the number of parameters (coefficients + intercept) whereas we use it for the number of independent variables. Readers should be aware of this difference.

where Y_i is the observed number of events for a zone i, P_i is the predicted number of events given a set of K independent variables, and $\bar{Y}$ is the mean number of events across zones. The R^2 is a number from 0 to 1; 0 indicates no predictability while 1 indicates perfect predictability.

R-square for OLS model

For an OLS model, R^2 is a very consistent estimate. It increases in a linear manner with predictability and is, therefore, a good indicator of how effective one model is compared to another. As with all diagnostic tests, the value of the R^2 increases with more independent variables. Consequently, R^2 is usually adjusted for degrees of freedom:

$$R_a^2 = 1 - \frac{\sum_{i=1}^{N} \frac{(Y_i - P_i)^2}{N-K-1}}{\sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1}} \tag{27.27}$$

where N is the sample size and K is the number of independent variables.

R-square for Poisson model

With the Poisson model, however, the R^2 value (whether adjusted or not) is not a good measure of overall fit. While the Poisson R^2 varies from 0 to 1, similar to the OLS, it is not monotonic. That is, the addition of a new variable to an equation often has unpredictable effects; sometimes it will increase substantially and sometimes it will increase only a little independent of how strong is a variable's association with the dependent variable (Miaou, 1996). This inconsistency comes from the decomposition of the total sum of squares:

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(Y_i - P_i)^2 + \sum_{i=1}^{N}(P_i - \bar{Y})^2 + 2\sum_{i=1}^{N}(Y_i - P_i)(P_i - \bar{Y}) \tag{27.28}$$

The first term in the equation is the residual sum of squares (or error term) while the second term is the explained sum of squares. In an OLS model, the third term is zero if an intercept is included (Cameron & Trivedi, 1998, 153). Hence, the total sum of squares is broken into two parts - that which is explained and that which is unexplained. However, for the Poisson and other non-linear regression methods, the last term is not zero. Consequently, a test that compares the explained sum of squares to the total sum of squares will not produce consistent results.

Other measures have been proposed, such as the deviance R-square which measures the reduction in the Likelihood Ratio due to the inclusion of predictor variables (Cameron & Windmeijer, 1996). It produces a slightly different R-square, one that is typically higher than the traditional R-square. Nevertheless, it has problems, too. Miaou (1996) argues that there is not a

single R-square index that is perfectly consistent. The AIC, BIC/SC and Deviance statistics (discussed in Chapter 16) are better indicators of goodness of fit.

Dispersion Parameter

Finally, in the Poisson model only, the dispersion parameter indicates the extent to which the variance is different from the mean. This was defined in equation 27.7 above.

Coefficients, Standard Errors, and Significance Tests

The second type of diagnostic test is those for the individual predictors in the model. In both the OLS and Poisson models, there are three tests:

1. The coefficient. This indicates the change in the dependent variable associated with the change in the independent variable. In the case of the OLS, it is a linear term (i.e., the value of the dependent variable is multiplied by the coefficient) while in the Poisson model, the change in the dependent variable is estimated by exponentiating the coefficient (i.e., $e^{\beta X}$).

2. The standard error. Each estimated coefficient in a model accounts for some of the variance in the dependent variable. This variance is the contribution of the particular independent variable to the variance of the dependent variable. The square root of that variance is the *standard error*.

3. The significance level. The ratio of the coefficient to the standard error produces a significance test of the coefficient. In the OLS model, it is a t-test with N-K-1 degrees of freedom whereas in the Poisson model it is an asymptotic t-test, which is effectively a Z-test. The appropriate tables (t-test or standard normal) produce approximate probability levels of a Type I error (the likelihood of falsely rejecting a true null hypothesis of no relationship).

Testing for Multicollinearity

One of the major problems with any regression model, whether OLS or Poisson, is multicollinearity among the independent variables. In theory, each independent variable should be statistically independent of the other independent variables. Thus, the amount of variance for the dependent variable that is accounted for by each independent variable should be a unique contribution. In practice, however, it is rare to obtain completely independent predictive variables. More likely, two or more of the independent variables will be correlated. The effect is that the estimated standard error of a predictor variable is no longer unique since it shares some of the

variance with other independent variables. The greater *communality* of shared variance, the more ambiguous will be the predicted effects. If two variables are highly correlated, it is not clear what contribution each makes towards predicting the dependent variable. In effect, multicollinearity means that variables are measuring the same effect.

Multicollinearity among the independent variables can produce very strange effects in a regression model. Among these effects are: 1) If two independent variables are highly correlated, but one is more correlated with the dependent variable than the other, the stronger one will usually have a correct sign while the weaker one will sometimes get flipped around (e.g., from positive to negative, or the reverse); 2) Two variables can cancel each other out; each coefficient is significant when it alone is included in a model but neither are significant when they are together; 3) One independent variable can inhibit the effect of another correlated independent variable so that the second variable is not significant when combined with the first one; and 4) If two independent variables are virtually perfectly correlated, many regression routines break down because the matrix cannot be inverted.

All these effects indicate that there is non-independence among the independent variables. Aside from producing confusing coefficients, multicollinearity can overstate the amount of predictability in a model. Since every independent variable accounts for some of the variance of the dependent variable, with multicollinearity the overall model will appear to improve when it probably has not.

Tolerance test

A user has to be aware of the problem of multicollinearity and seek to minimize it. The simplest solution is to drop variables that are co-linear with other independent variables already in the equation. A relatively simple test for assessing this is called *tolerance*. Tolerance is defined as *lack of predictability* of each independent variable by the other independent variables, or:

$$Tol_i = 1 - (R_{jk\ldots})^2 \qquad (27.29)$$

where $(R_{jk..})^2$ is the R^2 of an OLS equation where independent variable, i, is predicted by the other independent variables, j, k, l, and so forth. That is, *each* independent variable in turn is regressed against the *other* independent variables in the equation. The R^2 associated with that model is subtracted from 1. The higher the tolerance level, the less a particular independent variable shares its variance with the other independent variables.

Note that the tolerance test uses an OLS model; it assumes the dependent variable in the test (i.e., one of the independent variables) is normally distributed, which may or may not be true. Thus, in a Poisson or other non-linear model, one has to be careful about interpretation based on

the tolerance test. Nevertheless, the test can be a good indicator of whether two variables are collinear. As a rough guideline, a tolerance value of 0.7 or less usually indicates substantial multicollinearity. This level means that there is overlap of 50% or more in the variance of the tested variable with the other independent variables. A more strict and conservative approach uses a tolerance level of 0.8 or less as indicating multicollinearity. This level means that there is overlap of 36% or more in the variance of the tested variable with the other independent variables. An even stricter criterion is to use a tolerance level of 0.9 or less, essentially allowing 18% overlap in the variance of the tested variable with the other independent variables. In general, it is better to have a stricter model that has little multicollinearity. The interpretation of the coefficients will be cleaner and the model will generally be more reliable with other data sets.

Fixed model v. stepwise variable selection

There are several strategies designed to reduce multicollinearity in a model. One is to start with a defined model and eliminate those variables that have a low tolerance. The total model is estimated and the coefficients for each of the variables are estimated at the same time. This is sometimes called a *fixed model*. Then, variables that are co-linear are removed from the equation, and the model is re-run.

Another strategy is to estimate the coefficients a step at a time, a procedure known as *stepwise* regression (Der & Everitt, 2002, 88-89). There are several standard stepwise procedures. In the first procedure, variables are added one at a time (a *forward selection* model). The independent variable having the strongest linear correlation with the dependent variable is added first. Next, the independent variable from the remaining list of independent variables with the highest correlation with the dependent variable, *controlling for* the one variable already in the equation, is added next and the model is re-estimated. In each step, the independent variable with the highest correlation with the dependent variable controlling for the variables already in the equation is added to the model, and the model is re-estimated. This proceeds until either all the independent variables are added to the equation or else a stopping criterion is met. The usual criterion is only variables with a specified significance level are allowed to enter (called a *p-to-enter*).

A *backward elimination* procedure works in reverse. All independent variables are initially added to the equation. The variable with the weakest coefficient (as defined by the significance level) is removed, and the model is re-estimated. Next, the variable with the weakest coefficient in the second model is removed, and the model is re-estimated. This procedure is repeated until either there are no more independent variables left in the model or else a stopping criterion is met. The usual criterion is that all remaining variables pass a specified significance level (called a *p-to-remove*).

There are combinations of these procedures, for example adding variables in a forward selection but then removing any that are no longer significant or using a backward elimination procedure but allowing new variables to enter the model if they suddenly become significant.

There are advantages to each approach. A fixed model allows defined variables to be all included. If either theory or previous research has indicated that a particular combination of variables is important, then the fixed model allows that to be tested. A stepwise procedure might drop one of those variables. On the other hand, a stepwise procedure usually can obtain the same or higher predictability than a fixed procedure (whether predictability is measured by a log-likelihood or an R-square).

Within the stepwise procedures, there are also advantages and disadvantages to each method, though the differences are generally very small. A forward selection procedure adds variables one at a time. Thus, the contribution of each new variable can be seen. On the other hand, a variable that is significant at an early stage could become not significant at a later stage because of the unique combinations of variables. Similarly, a backward elimination procedure will ensure that all variables in the equation meet a specified significance level. But, the contribution of each variable is not easily seen other than through the coefficients. In practice, one usually obtains the same model with either procedure, so the differences are not that critical.

A stepwise procedure will not guarantee that multicollinearity will be removed entirely. However, it is a good procedure for narrowing down the variables to those that are significant. Then, any co-linear variables can be dropped manually and the model re-estimated. In the *CrimeStat* trip generation, both a fixed model and a backward elimination procedure are allowed.

Available Regression Models

CrimeStat has 10 different regression models that can be used for trip generation and which can be estimated with either maximum likelihood (MLE) or Markov Chain Monte Carlo (MCMC):

MLE Normal (OLS)
MCMC Normal
MCMC Normal-spatial autocorrelation component (CAR or SAR)
MLE Poisson
MLE Poisson with Linear Correction (NB1)
MLE Negative Binomial (Poisson-Gamma)
MCMC Poisson-Gamma
MCMC Poisson-Gamma-spatial autocorrelation component (CAR or SAR)
MCMC Poisson-Lognormal
MCMC Poisson-Lognormal-spatial autocorrelation component (CAR or SAR)

Users should consult Chapters 16 and 17 for details of these models. There are other methods for estimating the likely value of a count given a set of independent predictors. Among these are the zero-inflated Poisson (or ZIP; Hall, 2000), the Weibul function, the Cauchy function, and the lognormal function (see NIST 2004 for a list of common non-linear functions). There are also other spatial regression type models that correct for spatial autocorrelation in the dependent variable, such as geographically-weighted regression using a Poisson function (Fotheringham, Brunsdon, & Charlton, 2002). These are not included in this version of CrimeStat.

Adding Special Generators

In a travel demand model, there are *special generators*. These are unique land uses or environments that produce an extra large number of trips. For regular travel demand modeling, stadiums, airports, train stations, large parks, and mega-malls generate more than their share of trips, or at least than what would be predicted by the amount of permanent employment at those locations. They are usually attractors, not producers. In a normal transportation travel demand model, these zones are excluded from the cross-classification and independent estimates are made of them.

For crime trips, there are also special generators. Typically, these are zones that have more crimes being attracted to them than are expected on the basis of the population and employment at those locations. Since we are using a regression model to estimate the productions and attractions, a simple way to model a special generator is to create a simple *dummy* variable. This is a variable where zones with the special generator get a value ‘1' and zones without the special generator get a ‘0'. Essentially, the variable is a cross-classification of the special generator versus every other zone.

One has to be cautious is doing this, however. Typically, special generators are identified by having a greater number of crimes being attracted to a zone than is predicted by the model. In other words, they have a greater positive residual error (observed - predicted) and are ‘outliers’ in the residual error distribution. By adding a variable to explain those cases, the residual error decreases. But, in doing so, we are not really explaining why the zone has more crimes than expected, but simply accounting for it by putting in an empirical variable. In re-running the model, there will be, usually, new outliers that have a greater positive residual error. If this logic is to be repeated, then we would create new special generators for those zones and re-estimate the model. If continued without limits, eventually there would not be a model anymore but just a collection of dummy variables, one for each zone.

Therefore, a user should be cautious in introducing special generators. It is generally alright to introduce a few for the truly exceptional zones. These are zones where it is logical to treat them as special generators and where one would expect continuity over time. In other words, they

should be used if the special generator status is expected to last over time. For example, a stadium or an airport or a train station is liable to remain at its location for many years. A particular shopping mall, on the other hand, may attract crimes at one particular point in time but not necessarily in the future. Unless a mall is so much larger than other malls in the region (a mega-mall), it should not be assigned a special generator status.

Adding External Trips

External trips are, by definition, trips that come from outside the region. They are part of the origin/production model in that these are trips that are not accounted for by the model. There are also trips that originate within the study area, but end outside the area; however, those are usually not modeled since the focus will be on the study area itself. In the usual travel demand framework, external trips are those coming from major corridors into the region. Estimates of the travel on these corridors are obtained by *cordon counts*, counts of vehicles coming into the region and leaving the region (net inflow). Estimates of future growth of those external trips has to based on expectations of future population growth the metropolitan region and in nearby regions.

For crime trips, external trips are defined as trips that originate outside the study area. But they must be estimated by the difference between the total number of crimes occurring in the destination study area and the total originating in the origin zones. That is, of all the crimes occurring in the study area, the origin zones are modeled. Those trips that originate from outside the origin zones are external trips. They must be added to the predicted number of origin trips to produce an adjusted estimate of total origins, or:

$$O_j = O_{pi} + O_e \tag{27.30}$$

where O_j is the total number of crime origins for crimes committed in study area, j, O_{pi} is the total number of crimes originating in the origin zones, i, and O_e is the total number of crimes originating outside the region, e.

In other words, for the production (origin) model **only**, we add an external zone to account for crime trips that originated outside the modeled region. If that is not done, in the balancing step the number of crimes originating in each zone will be overestimated because the predicted origins will be multiplied by a factor to ensure that the total number of origins equals the total number of destinations.

Not including the external trips can lead to bias in the model. If the number of external trips is a sizeable percentage of all crime origins occurring in the study area, then the coefficients of the origin model could be misleading. In practice, most travel demand modelers assume that if

the percentage of external trips is not greater than 5%, there usually is little bias introduced (Ortuzar & Willumsen, 2001). If it is greater than 5%, then origin zones from adjacent jurisdictions need to be included in the origin model.

Balancing Predicted Origins and Predicted Destinations

The trip generation 'model' is actually two separate models: 1) a model of trips produced by every zone and 2) a model of trips attracted to every zone. Since a trip has an origin and a destination (by definition), then the total number of productions must equal the total number of attractions,

$$\sum_{i=1}^{M} O_i = \sum_{j=1}^{N} D_j \qquad (27.31)$$

where *O* is a trip origin, *D* is a trip destination, and i and j are zone numbers. Note that in equation 27.31, there are *M* origin zones and *N* destination zones. This implies that *M* and *N* do not have to be equal. In fact, including an external zone guarantees that *M* and *N* will not be equal (*M* will be at least 1 greater than *N*). If an entire metropolitan area is being modeled, the *M* and *N* will be almost identical (differing only in the external zone). However, if the study area being modeled is a sub-set of the metropolitan region, then *M* will be much greater than *N*. For example, in modeling crime trips in Baltimore County, there are 532 origin zones (including those from Baltimore County and from the City of Baltimore) and only 325 destination zones (only those in Baltimore County).

To ensure that this equality is true, a balancing operation is conducted. Essentially, this means multiplying either the number of predicted origins in each origin zone or the number of predicted destinations in each destination zone by a constant which is the ratio of either the total destinations to the total origins (to multiply the number of predicted origins) or the ratio of the total origins to the total destinations (to multiply the number of predicted destinations).

With crime analysis, the number of destinations would generally be considered a more reliable data set than the number of origins. Because crimes are enumerated where they occur, the number of crimes occurring at any one location is more accurate than the location of the offenders. Thus, we adjust the predicted origins so that they equal the predicted destinations.[8]

[8] In the usual travel demand modeling, on the other hand, modelers usually adjust the predicted destinations since the origin data is more reliable. These numbers are obtained from the census or from the sample of households who are interviewed to produce a sample from which data on destinations are obtained.

Summary of the Trip Generation Model

In summary, the trip generation model is estimated in four steps:

1. A model of the predictors of the number of crimes origins (a crime production model);
2. A model of predictors of the number of crime destinations (a crime attraction model);
3. External trips are estimated and added to the number of predicted origins as an external zone; and
4. The total number of predicted crime origins is balanced to be equal to the total number of predicted crime destinations.

The *CrimeStat* Trip Generation Model

In this section, we describe the trip generation model implemented in *CrimeStat*. As mentioned above, this step involves calibrating a regression model against the zonal data. Two separate models are developed, one for trip origins and one for trip destinations. The dependent variable is the number of crimes originating in a zone (for the trip origin model) or the number of crimes ending in a zone (for the trip destination model). The independent variables are zonal variables that may predict the number of origins or destinations.

There are three steps to the model, each corresponding to a separate tab in *CrimeStat*:

1. Calibrate the model
2. Make a prediction
3. Balance the predicted origins and the predicted destinations

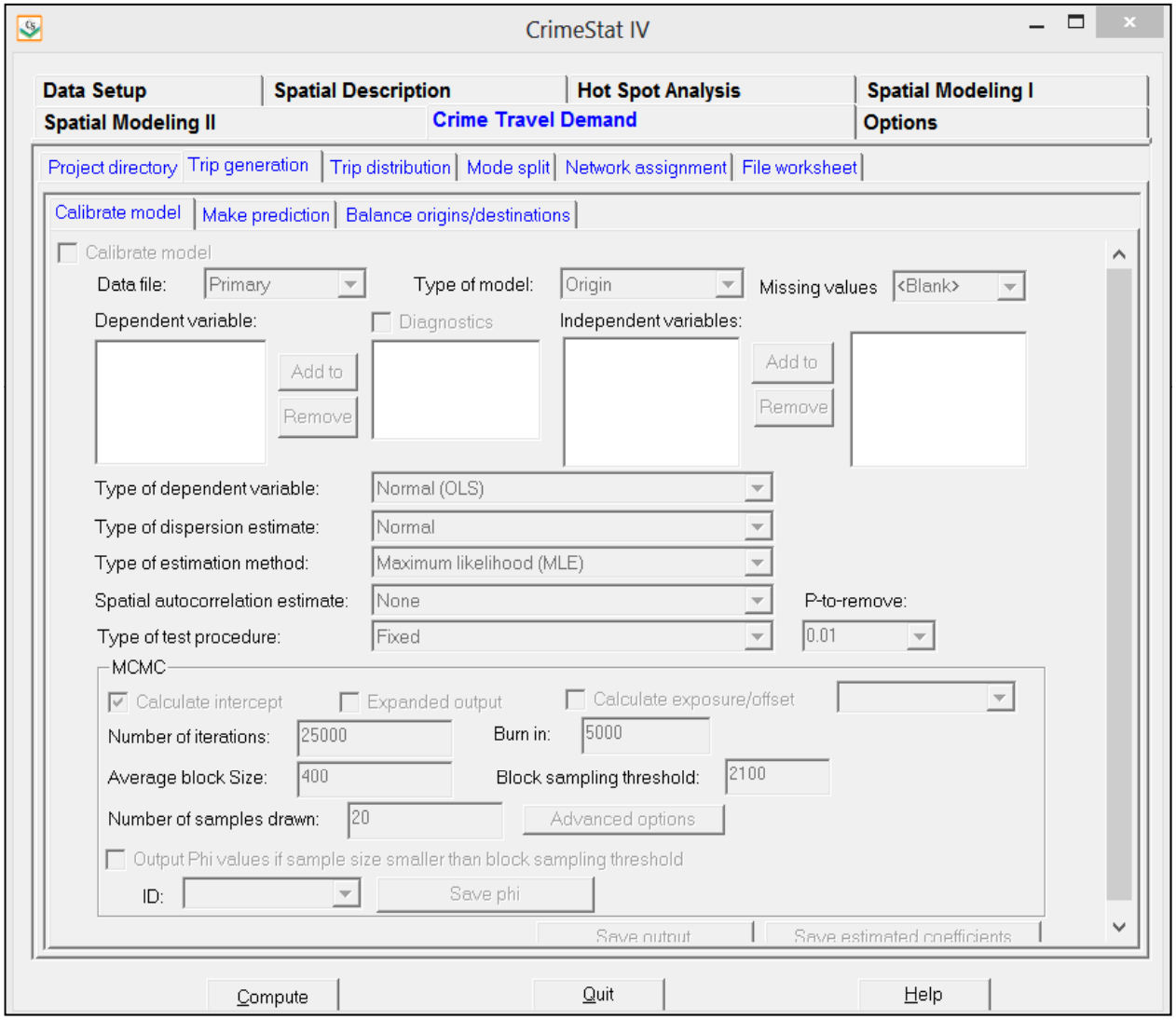
Figure 27.5 shows an image of the trip generation model page within *CrimeStat*. The trip generation model is made up of three separate pages (or tabs):

1. A *Calibrate model* page in which a regression model can be run to estimate either an origin (production) model or a destination (attraction) model;

2. A *Make prediction* page in which the estimated coefficients can be applied to the same or a different data set and in which the external trips can be added to the predicted origins; and

Figure 27.5:

Trip Generation Module



3. A *Balance predicted origins & destinations* page in which the total predicted origins can be adjusted to equal the total predicted destinations.

Calibrate Model

In the first step, models are calibrated using the input data. There is a model for the origin zones and another model for the destination zones. The user should indicate what type of model is being run in order to make the output more clear.

Data File

The data file is input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Type of Model

Specify whether the model is for origins or destinations. This will be printed out on the output header.

Dependent Variable

Select the dependent variable from the list of variables. There can be only one dependent variable per model.

Skewness Diagnostics

If checked, the routine will test for the skewness of the dependent variable. The output includes:

1. The 'g'statistic
2. The standard error of the 'g' statistic
3. The Z value for the 'g' statistic
4. The probability level of a Type I error for the 'g' statistic
5. The ratio of the sample variance to the sample mean

Error messages indicate whether there is probable skewness in the dependent variable. If there is skewness, use a Poisson regression model.

Independent variables

Select independent variables from the list of variables in the data file. Up to 15 variables can be selected.

Missing values

Specify any missing value codes for the variables. Blank records will automatically be considered as missing. If any of the selected dependent or independent variables have missing values, those records will be excluded from the analysis.

Type of Regression Model

Specify the type of regression model to be used. The default is a Poisson regression with over-dispersion correction (NB1). Other alternatives are:

1. Ordinary Least Squares regression;
2. Poisson regression;
3. MLE Poisson-Gamma (negative binomial; NB2);
4. MCMC Poisson-Gamma;
5. MCMC Poisson-Lognormal; and
6. MCMC Conway-Maxwell Poisson.

Each of the MCMC models can be run with a spatial autocorrelation component added, either a CAR or a SAR. See Chapters 16 and 17 for more details.

Type of Regression Procedure

If the model being run is an MLE routine (Poisson, Poisson with linear correction (NB1), or Poisson-Gamma (NB2), specify whether a fixed model (all selected independent variables are used in the regression) or a backward elimination stepwise model is used. The default is a fixed model. If a backward elimination stepwise model is selected, choose the P-to-remove value (default is .01). The backward elimination starts with all selected variables in the model (the fixed procedure). However, it proceeds to drop variables that fail the P-to-remove test, one at a time. Any variable that has a significance level in excess of the P-to-remove value is dropped from the equation.

With MCMC routines, however, only fixed models can be run.

Save Estimated Coefficients/Parameters

The estimated coefficients of the final model can be saved as a 'dbf' file. Specify a file name. This would be useful in order to repeat the regression while adding in external trips to the predicted origins (see Make trip generation prediction below) or to apply the coefficients to another dataset (e.g., future values of the independent variable).

Save Output

The output is saved as a 'dbf' file under a different file name. The output includes all the variables in the input data set plus two new ones: 1) the predicted values of the dependent variable for each observation (with the name PREDICTED); and 2) the residual error values, representing the difference between the actual /observed values for each observation and the predicted values (with the name RESIDUAL).

Poisson output

The output of the Poisson regression routines includes 13 fields for the entire model:

1. The dependent variable
2. The type of model
3. The sample size (N)
4. The degrees of freedom (N - # dependent variables – 1)
5. The type of regression model (Poisson, Poisson with over-dispersion correction)
6. The log-likelihood value
7. The Likelihood Ratio
8. The probability value of the Likelihood Ratio
9. The Akaike Information Criterion (AIC)
10. The Bayesian Information Criterion/Schwartz Criterion (BIC/SC)
11. The Dispersion Multiplier
12. The approximate R-square value
13. The deviance R-square value

and 5 fields for each estimated coefficient:

14. The estimated coefficient
15. The standard error of the coefficient
16. The pseudo-tolerance value of the coefficient (see below)

17. The Z-value of the coefficient
18. The p-value of the coefficient.

OLS output

The output of the Ordinary Least Square (OLS) routine includes 9 fields for the entire model:

1. The dependent variable
2. The type of model
3. The sample size (N)
4. The degrees of freedom (N - # dependent variables – 1)
5. The type of regression model (Norma/Ordinary Least Squares)
6. Squared multiple R
7. Adjusted squared multiple R
8. F test of the model
9. p-value of the model

and 5 fields for each estimated coefficient:

10. The estimated coefficient
11. The standard error of the coefficient
12. The tolerance value of the coefficient (see below)
13. The t-value of the coefficient
14. The p-value of the coefficient.

Multicollinearity Among Independent Variables

To test multicollinearity, a tolerance test is run (see equation 27.29 above). There is not a simple test of whether a particular tolerance is meaningful or not. In *CrimeStat*, several qualitative categories are used and error messages are output:

1. If the tolerance value is 0.80 or greater, then there is little multicollinearity (No apparent multicollinearity);

2. If the tolerance is between 0.60-0.79, there is some multicollinearity (possible multicollinearity);

3. If the tolerance is between 0.25-0.59, there is probable multicollinearity (probable multicollinearity. Eliminate variable with lowest tolerance and re-run); and

4. If tolerance is less than 0.25, there is definite multicollinearity (Definite multicollinearity. Results are not reliable. Eliminate variable with lowest tolerance and re-run).

Graph

While the output page is open, clicking on the graph button will display a graph of the residual errors (on the Y axis) against the predicted values (on the X axis).

Make Trip Generation Prediction

This routine applies an already-calibrated regression model to a data set. This would be useful for several reasons: 1) if external trips are to be added to the model (which is normally preferred); 2) if the model is applied to another data set; and 3) if variations on the coefficients are being tested with the same data set. The model will need to be calibrated first (see Calibrate Trip Generation Model) and the coefficients saved as a parameters file. The coefficient parameter file is then re-loaded and applied to the data.

Data File

The data file is input as either the primary or secondary file. Specify whether the data file is the primary or secondary file.

Type of Model

Specify whether the model is for origins or destinations. This will be printed out on the output header.

Trip Generation Coefficients/Parameters File

This is the saved coefficient parameter file. It is an ASCII file and can be edited if alternative coefficients are being tested (be careful about editing this without making a backup). Load the file by clicking on the Browse button and finding the file. Once loaded, the variable names of the saved coefficients are displayed in the 'Matching parameters' box.

Independent Variables

Select independent variables from the list of variables in the data file. Up to 15 variables can be selected.

Matching parameters

The selected independent variables need to be matched to the saved variables in the trip generation parameters file in the same order. Add the appropriate variables one by one in the order in which they are listed in the matching parameters box. It is essential that the order by the same otherwise the coefficients will be applied to the wrong variables.

> Hint: With your cursor placed in the list of independent variables, typing the first letter of the matching variable name will take you to the first variable that starts with that letter. Repeating the letter will move down the list to the second, third, and so forth until the desired variable is reached.

Specify any missing value codes for the variables. Blank records will automatically be considered as missing. If any of the selected dependent or independent variables have blank values, those records will be excluded from the analysis.

Add External Trips

External trips are those that start outside the modeled study area. Because they are crimes that originate outside the study area, they were not included in the zones used for the origin model. Therefore, they have to be independently estimated and added to the origin zone total to make the number of origins equal to the number of destinations. Click on the 'Add external trips' button to enable this feature.

Origin ID

Specify the origin ID variable in the data file. The external trips will be added as an extra origin zone, called the 'External' zone. Note: the ID's used for the destination file zones should be the same as in the origin file. This will be necessary in subsequent modeling stages.

Number of external trips

Add the number of external trips to the box. This number will be added as an extra origin zone (the External zone).

Type of Regression Model

Specify the type of regression model to be used. The default is a Poisson regression and the other alternative is an Ordinary Least Squares regression.

Save Predicted Values

The output is saved as a 'dbf' file under a different file name. The output includes all the variables in the input data set plus the predicted values of the dependent variable for each observation (with the name PREDICTED). In addition, if external trips are added, then there should be a new record with the name EXTERNAL listed in the Origin ID column. This record lists the added trips in the PREDICTED column and zeros (0) for all other numeric fields.

Output

The tabular output includes summary information about file and lists the predicted values for each input zone.

Balance Predicted Origins & Destinations

Since, by definition, a 'trip' has an origin and a destination, the number of predicted origins must equal the number of predicted destinations. Because of slight differences in the data sets of the origin model and the destination model, it is possible that the total number of predicted origins (including any external trips) may not equal the total number of predicted destinations. This step, therefore, is essential to guarantee that this condition will be true. The routine adjusts either the number of predicted origins or the number of predicted destinations so that the condition holds. The trip distribution routines will not work unless the number of predicted origins equals the number of predicted destinations (within a very small rounding-off error).

Predicted Origin File

Specify the name of the predicted origin file by clicking on the Browse button and locating the file.

Origin variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED).

Predicted Destination File

Specify the name of the predicted destination file by clicking on the Browse button and locating the file.

Destination variable

Specify the name of the variable for the predicted origins (e.g., PREDICTED).

Balancing Method

Specify whether origins or destinations are to be held constant. The default is 'Hold destinations constant'.

Save Predicted Origin/Destination File

The output is saved as a 'dbf' file under a different file name. The output includes all the variables in the input data set plus the adjusted values of the predicted values of the dependent variable for each observation. If destinations are held constant, the adjusted variable name for the predicted trips is ADJORIGIN. If origins are held constant, the adjusted variable name for the predicted trips is ADJDEST.

Output

The tabular output includes file summary information plus information about the number of origins and destinations before and after balancing. In addition, the predicted values of the dependent variable are displayed.

Example of the Trip Generation Model

To illustrate this model, an example from Baltimore County. In the case of Baltimore County, MD, will be used. The zonal geography is traffic analysis zones (TAZ). Two data sets were produced, one for the crime origins and one for the crime destinations. For Baltimore County, the origin data set had 532 zones covering both Baltimore County and the City of Baltimore with the total number of crime origins for each zone (sub-divided into different crime types - robberies, burglaries, vehicle theft) and a number of possible predictor variables (population, retail and non-retail employment, median household income, poverty levels, and vehicle ownership). Similarly, the destination data set had 325 zones with the number of crime destinations for each zone (again, sub-divided into different crime types) and number of possible predictor variables (population, retail and non-retail employment, median household income, and

several land use categories - acreage allocated for retail, residential, office space, and conservation uses). Sample data sets are provided on the *CrimeStat* download page.

Setting Up the Origin Model

In the first step, an origin model is created. Figure 27.6 shows the selection of the dependent variable and some possible independent variables. The type of model is an ordinary Poisson regression. The dependent variable is the number of crimes occurring between 1993 and 1997 in each origin zone (BCORIG). Eight possible independent variables have been selected:

1. 1996 population of each zone (POPULATION)
2. 1990 median household income of the zone relative to the zone with the highest median household income (INCOME EQUALITY)
3. Number of 1996 non-retail employees in each zone (NON-RETAIL EMPLOYMENT)
4. Number of 1996 retail employees in each zone (RETAIL EMPLOYMENT)
5. Total linear miles of arterial roads in each zone (ARTERIAL ROADS)
6. A dummy variable for whether the Baltimore Beltway (I-695) passed through the zone or not (BELTWAY)
7. Linear distance of the zone from Baltimore harbor in the CBD (DISTANCE FROM CENTER); and
8. 1990 Number of households without automobiles (HOUSEHOLDS WITH NO AUTOMOBILES)

The model is set up to run an ordinary Poisson regression (without an adjustment to the dispersion). It is a fixed model in which all independent variables are included. The coefficients are saved under 'Save estimated coefficients' dialogue box and the output (the predicted values) are saved under the 'Save output' dialogue box. Both boxes ask for a file name.

Table 27.2 shows the results. The format is simplified from that shown in Chapter 16. Key statistics are highlighted. The overall model is highly significant. The log likelihood is shown as are the AIC and BIC/SC adjusted log likelihood. The deviance and Pearson are highly significant, indicating that the model predicts significantly better than chance. The coefficients for each of the variables are all significant.

However, there are two major problems. First, the dispersion multiplier (parameter) is very large (36.09) and significant, indicating that the conditional variance is more than 36 times greater than the conditional mean. Second, while all of the coefficients are significant, several show sizeable multicollinearity as evidenced by the pseudo-tolerance value (POPULATION, DISTANCE, HOUSEHOLDS WITH NO AUTOMOBILES as well as INCOME EQUALITY). This indicates that these variables are essentially measuring the same thing.

Table 27.2:
Full Origin Model: Poisson

Model result:		
Data file:	BaltOrigins.dbf	
Type of model:	Origin	
DepVar:	**BCORIG**	
N:	532	
Df:	522	
Type of regression model:	MLE Poisson	
Likelihood statistics		
Log Likelihood:	-10,678.05	
AIC:	21,376.10	
BIC/SC:	21,418.87	
Deviance:	18,547.38	p≤.0001
Pearson Chi-square:	19,396.48	p≤.0001
Model error estimates		
Mean absolute deviation:	38.70	
Mean squared predicted error:	3,920.66	
Dispersion tests		
Dispersion multiplier:	36.09	p≤.01

Predictor	**Coefficient**	**Stand Error**	**Tolerance**	**Z-value**	**p-value**
CONSTANT	**4.1890**	0.0202	.	207.03	0.001
POPULATION	**0.0003**	0.000003	**0.46**	121.13	0.001
INCOME EQUALITY	**-0.0330**	0.0007	**0.61**	-48.85	0.001
NON-RETAIL EMPLOYMENT	**-0.0002**	0.000005	0.84	-36.87	0.001
RETAIL EMPLOYMENT	**-0.0004**	0.00002	0.96	-18.91	0.010
ARTERIAL ROAD	**-0.1083**	0.0059	0.77	-18.49	0.001
BELTWAY	**0.1510**	0.0193	0.96	7.84	0.001
DISTANCE FROM CENTER	**0.0343**	0.0016	**0.49**	21.09	0.001
HOUSEHOLDS WITH NO AUTOMOBILES	**-0.0005**	0.00002	**0.36**	-18.95	0.010

Restructuring the Origin Model

Consequently, the model was restructured in three ways (Figure 27.7). First, to correct for over-dispersion, an MLE Poisson-Gamma (negative binomial) model was run. This is the most common approach to handling over-dispersion (see Chapter 16). Second, two co-linear variables - DISTANCE and ZEROAUTO, were dropped from the model. Third, a stepwise backward elimination procedure is used with the probability for keeping a variable in the equation (p-to-remove) being 0.01; that is, unless the probability that a coefficient could be obtained by chance is less than 1 in 100, the variable was dropped.

Table 27.3:
Reduced Origin Model: Poisson-Gamma

Model result:		
Data file:	BaltOrigins.dbf	
Type of model:	Origin	
DepVar:	**BCORIG**	
N:	532	
Df:	526	
Type of regression model:	MLE Poisson-Gamma	
Likelihood statistics		
Log Likelihood:	-2,627.65	
AIC:	5,267.30	
BIC/SC:	5,292.96	
Deviance:	623.49	p≤.0001
Pearson Chi-square:	500.59	p≤.0001
Model error estimates		
Mean absolute deviation:	57.28	
Mean squared predicted error:	18,143.78	
Dispersion tests		
Dispersion multiplier:	0.74	n.a.

Predictor	**Coefficient**	**Stand Error**	**Tolerance**	**Z-value**	**p-value**
CONSTANT	**3.4832**	0.131	.	26.61	0.001
POPULATION	**0.0004**	0.00003	0.95	17.10	0.001
INCOME EQUALITY	**-0.0178**	0.003	0.91	-5.46	0.001
NON-RETAIL EMPLOYMENT	**-0.0001**	0.00002	0.87	-6.71	0.001
RETAIL EMPLOYMENT	**-0.0002**	0.0001	0.96	-2.17	0.05

Figure 27.6:

Origin Poisson Model Setup

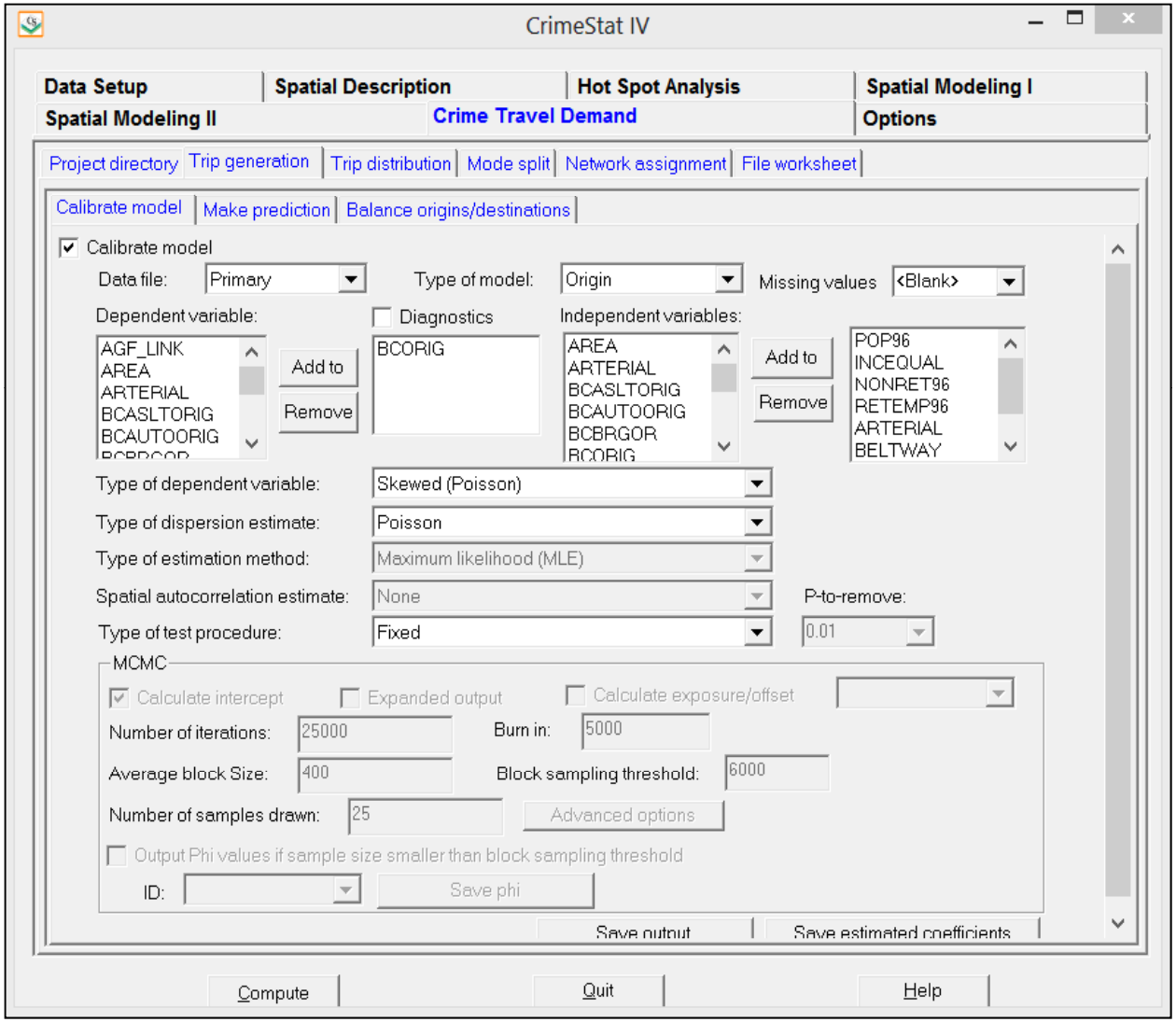
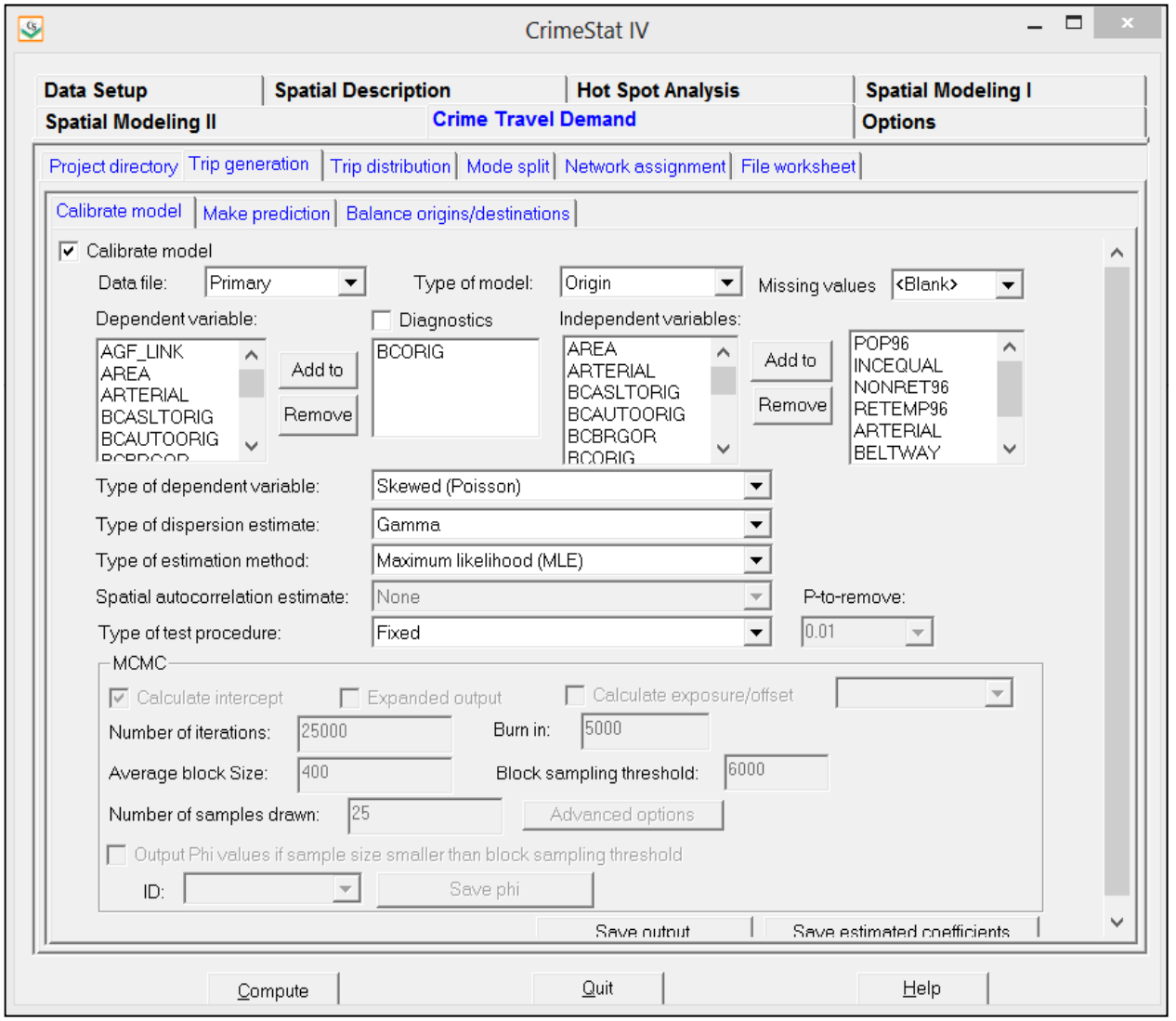


Figure 27.7:

Origin Poisson-Gamma Model Setup



The result is a model with four significant variables. Note that the full Poisson model (Table 27.3) has a greater negative log likelihood and a much greater AIC and BIC/SC value than the reduced model. This is because the reduced model was tested with a Poisson-Gamma mixed function and has a different probability structure. To properly compare it, the full model was run as a Poisson-Gamma (not shown). In the reduced model, the log likelihood was -2,627, which is even stronger than -2,618 for the full model, while the AIC was 5,267 and the BIC/SC was 5,293, compared to 5,526 and 5,299 for the full model respectively. In other words, the reduced model produced likelihood values very similar to the full model. More importantly, the overall fit of the model was almost as good as the full model. The mean absolute deviation was 57, compared to 55 for the full model, while the mean squared predicted error was 18,144, compared to 17,881 for the full model. Most importantly, the dispersion parameter is now less than 1.0.[9]

In other words, we have a simpler model that predicts almost as well as the full model but with coefficients that are less ambiguous. Such a model is liable to be more stable because the Poisson-Gamma has adjusted for over-dispersion in the Poisson while collinear and less significant variables have been removed.

Looking at the model, we see four variables that significantly predict the number of crime origins. Population is the strongest, as indicated by its Z-test. Non-retail employment is the next strongest with a negative coefficient (i.e., zones with less non-retail employment generate more crime trips). This is followed by relative income equality is the next strongest, also with a negative coefficient (i.e., zones with low relative income equality produce more crime origins). The fourth variable is retail employment and, like non-retail employment, the coefficient is also negative. In other words, zones with less overall employment produce more crime trips.

Residual Analysis of Origin Model

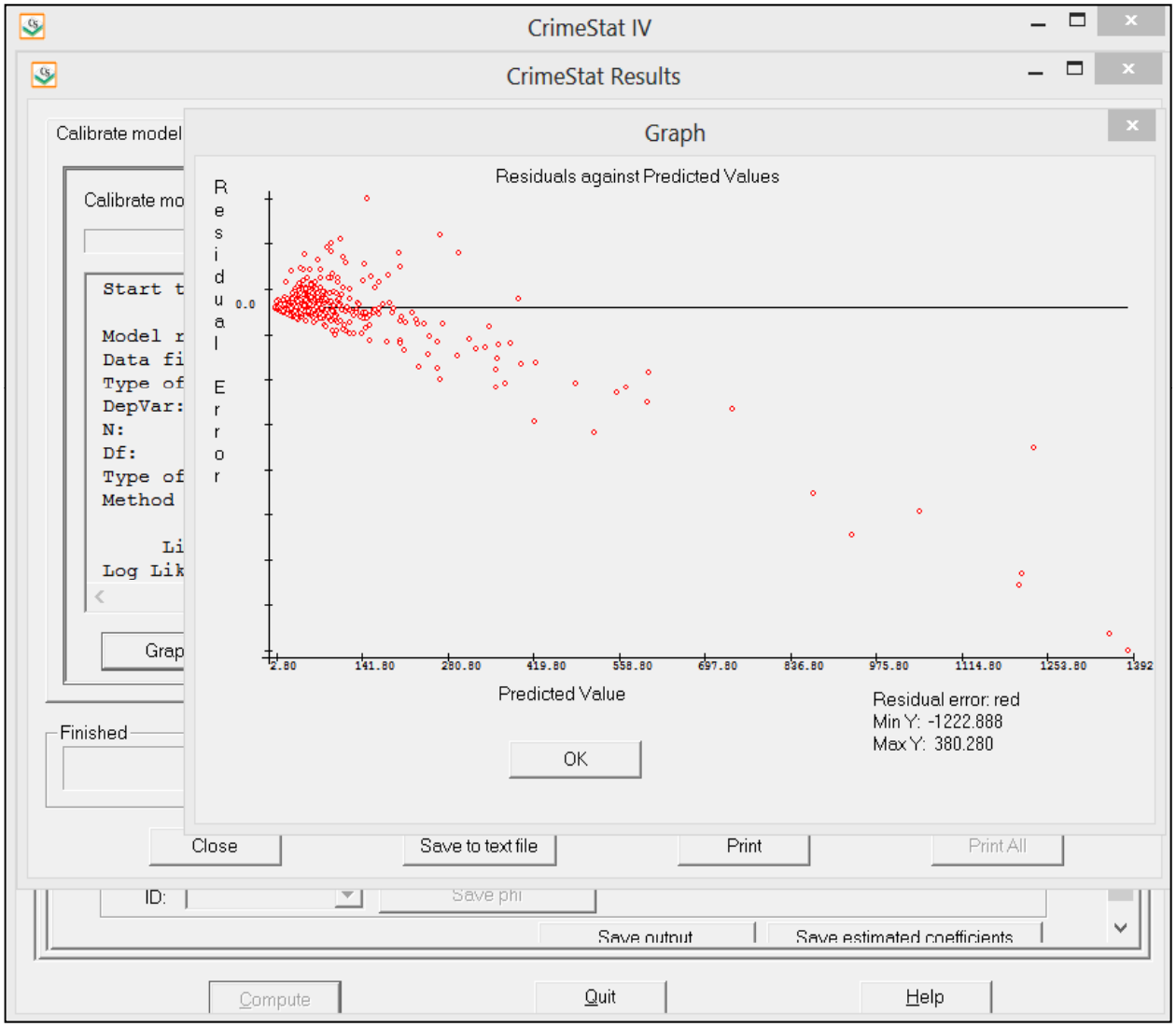
The *CrimeStat* output includes a graph of the residual errors (actual values minus the predicted values) on the Y-axis by the predicted values on the X-axis. It is important to examine the residual errors as these can indicate outliers, problems in the data, and violation of assumptions. Figure 27.8 shows an image of the residual graph screen. As seen, the errors increase with the value of the predicted dependent variable. With the Poisson model, this is expected and does not indicate the violation of the independent errors assumption, as it does with the OLS. The errors are reasonably symmetrical and do not indicate differences in over- and under-estimation across the band of the predicted values.

There are some outliers. There are two zones in which the predicted number of crimes originating from the zones substantially exceeded the number that actually originated from those

[9] A test of the dispersion parameter is not appropriate since it only tests for over-dispersion, not under-dispersion.

Figure 27.8:

Plot of Residual Errors and Predicted Values



zones and there is one zone that had more crimes originate from it than was predicted by the model. But, in general, the model appears to be reasonably balanced.

Setting Up the Destination Model

The same logic was applied for the destination model. In this case, the destination file has data on 325 zones within Baltimore County only. Similar possible predictor variables are included in the file. Aside from population, retail and non-retail employment, and the roadway variables, more detailed analysis on land uses were included (acreage of commercial, residential, office space, recreational, and conservation lands). The model that was run was a Poisson-Gamma (negative binomial) because the simple Poisson showed very high over-dispersion. Again, a backward elimination procedure was adopted. Once a final model was selected, it was re-run as a fixed model to ensure that the coefficients were consistently estimated. Table 27.4 presents the results.

Four variables ended up in the final model. Again, population was significantly related to the number of crimes attracted to a zone, but was not the strongest predictor as indicated by the Z-test. The strongest relationship was for the number of retail employees. This suggests that retail/commercial areas attract many crimes. Two other variables are in the equation. Relative income equality was, again, negatively related to crime destinations/attractions; zones with low income tend to attract more crimes. Also, there was a negative association with distance from the CBD. The farther away from the CBD was the zone, the lower the number of crimes. Overall, the model suggests that zones with commercial activities, which are closer to the city center, and which have households with relatively lower incomes are those that attract the most crimes.

The overall model was highly significant, as indicated by the Deviance and the Pearson Chi-square. The amount of multicollinearity is very low, which is ideal. Even though a model with more negative log likelihood (and more positive AIC and BIC/SC) could be produced by adding more variables, the amount of multicollinearity would be substantial. The philosophy expressed here is that a simpler model, but with little multicollinearity, is to be preferred over a more complex model but where the coefficients are less stable and more ambiguous. Generally, simpler models hold up better with new data sets (Radford, 2006; Nannen, 2003).

Residual Analysis of Destination Model

As with the origin model, an analysis was conducted of the residual errors. This time, the output 'dbf' file was brought into Excel and a nicer graph created (Figure 27.9). Unlike the best origin model, the dispersion of the residuals is not symmetrical. There are several major outliers, both on the negative end of the residuals (over-estimation of crime attractions) and on the positive end (under-estimation of crime attractions). In particular, there are two zones that seem to stand

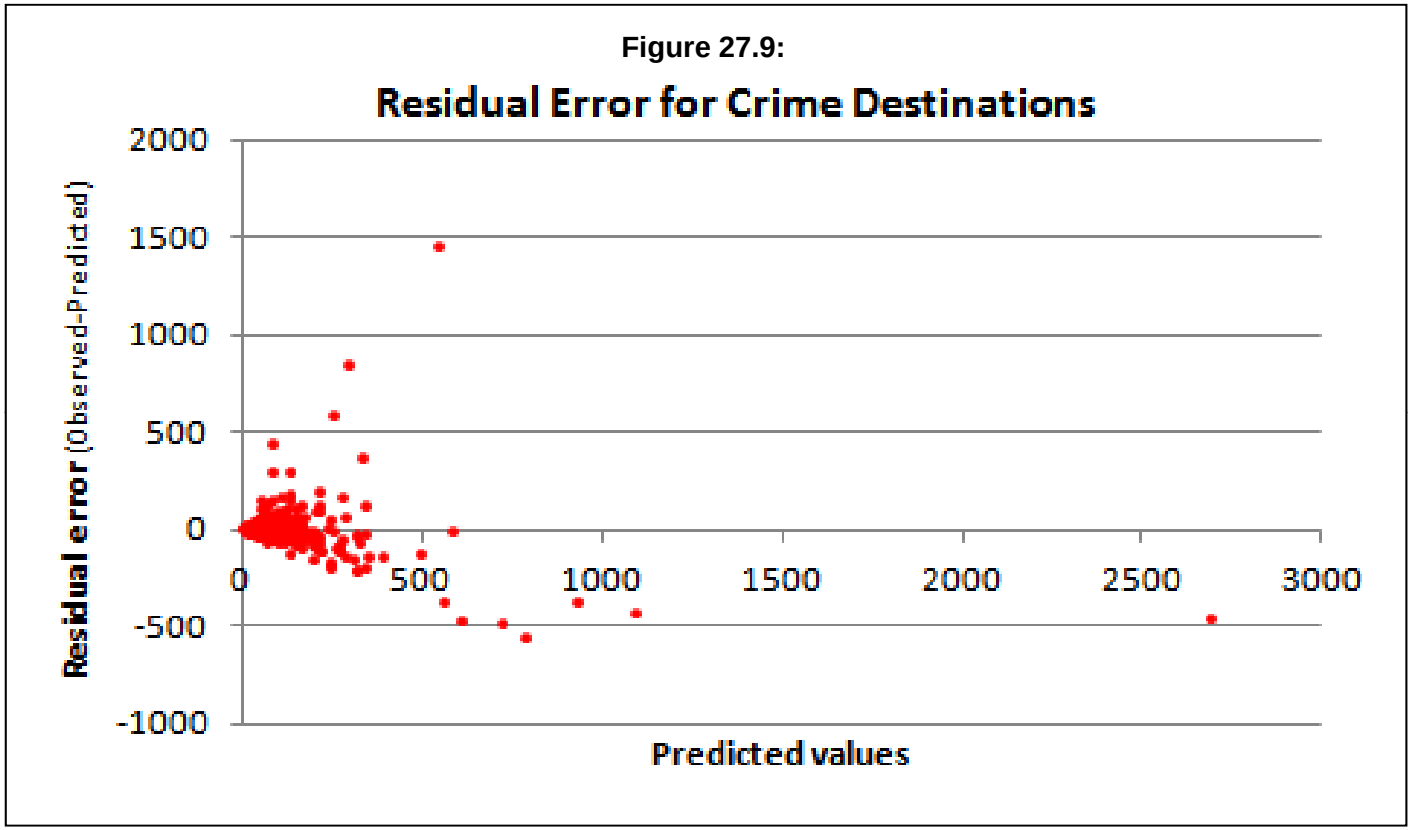
out. Both of them have shopping malls (Golden Ring Mall and Eastpoint Mall) but the amount of crime in those zones was much greater than the model predicted. This is seen as high positive residuals (i.e., there were more actual crimes than predicted). They both are older malls, but are located in relatively high crime areas. Golden Ring Mall was demolished some years ago, but after the data used in this example were collected.

Table 27.4:
Reduced Destination Model: Poisson-Gamma

Model result:		
Data file:	BaltOrigins.dbf	
Type of model:	Origin	
DepVar:	**BCDEST**	
N:	325	
Df:	319	
Type of regression model:	MLE Poisson-Gamma	
Likelihood statistics		
Log Likelihood:	-1,697.01	
AIC:	3,406.03	
BIC/SC:	3,428.73	
Deviance:	350.74	p≤.0001
Pearson Chi-square:	379.87	p≤.0001
Model error estimates		
Mean absolute deviation:	167.43	
Mean squared predicted error:	2,893,931.60	
Dispersion tests		
Dispersion multiplier:	0.43	n.a.

Predictor	**Coefficient**	**Stand Error**	**Tolerance**	**Z-value**	**p-value**
CONSTANT	**4.5208**	0.153	.	29.53	0.001
POPULATION	**0.0003**	0.00003	0.94	12.06	0.001
INCOME EQUALITY	**-0.0213**	0.003	0.90	-7.87	0.001
RETAIL EMPLOYMENT	**0.0020**	0.0001	0.94	17.49	0.001
DISTANCE FROM CENTER	**-0.0714**	0.010	0.88	-7.46	0.001

Figure 27.9:


Residual Error for Crime Destinations
Residual error (Observed-Predicted)
2000
1500
1000
500
0
-500
-1000
0
500
1000
1500
2000
2500
3000
Predicted values

Adding in Special Generators

Since the number of crime incidents (attractions) in those two zones was much higher than expected, they were treated as 'special generators'. Keeping in mind the caution that one does not want to over-use this category, a demonstration of how it works will be illustrated. Two new variables were created for the data set. One was for the Golden Ring Mall and one was for the Eastpoint Mall. For the Golden Ring Mall, the zone that included it received a '1' for this variable while all other zones received a '0'. Similarly, for the Eastpoint Mall variable, the zone in which it occurred received a '1' while all other zones received a '0'. These *dummy* variables were then included in the model (Table 27.5).

Adding the two special generators produces a model that, on the face of it, has not improved the predictability.. The log likelihood value is less negative than without the special generators and the AIC and BIC/SC statistics are also lower. The deviance values are about the same.

However, the Pearson Chi-square is quite a bit higher with the special generators. Also, the Mean Absolute Deviation (MAD) and the Mean Squared Predictive Error (MSPE) are substantially better with the special generators. This indicates that the new model which includes the special generators fit the data much better.

The coefficients for the two zones, treated as special generators, are both significant though not as strongly as the other variables. All other variables have the same relationships as in the first run. There is little multicollinearity. In other words, adding dummy variables for the two zones with higher than expected numbers of crime committed has produced a closer fitting model than not including the dummy variables.

This brings up an issue over the status of a special generator. In this example, the two zones were treated as special generators in the model. While the model fit increased substantially, one has to wonder whether this was a meaningful operation or not? That is, if this model were applied to data for a later time period (e.g., 2010-2012 crime data), would the relationships still hold? In the case of the Golden Ring Mall, it would not since that mall has since been demolished

The value of a special generator is that it identifies a land use that would be expected to be relatively permanent (e.g., a stadium or a train station or an airport). If it is a high visibility 'regional' mall, then treating it as a special generator is probably a good idea. If it is a smaller, older mall, on the other hand, the analysis is guessing that the mall will maintain its status as a high crime attraction location. Clearly, judgment and knowledge of the particular mall is essential.

Table 27.5:

Destination Model with Special Generators: Poisson-Gamma

Model result:		
Data file:	BaltOrigins.dbf	
Type of model:	Origin	
DepVar:	**BCDEST**	
N:	325	
Df:	319	
Type of regression model:	MLE Poisson-Gamma	
Likelihood statistics		
Log Likelihood:	-1,688.30	
AIC:	3,392.60	
BIC/SC:	3,422.87	
Deviance:	350.89	p≤.0001
Pearson Chi-square:	398.73	p≤.0001
Model error estimates		
Mean absolute deviation:	116.54	
Mean squared predicted error:	1,017,064.23	
Dispersion tests		
Dispersion multiplier:	0.40	n.a.

Predictor	**Coefficient**	**Stand Error**	**Tolerance**	**Z-value**	**p-value**
CONSTANT	**4.4625**	0.149	.	29.91	0.001
POPULATION	**0.0004**	0.00003	0.93	12.72	0.001
INCOME EQUALITY	**-0.0205**	0.003	0.90	-7.84	0.001
RETAIL EMPLOYMENT	**0.0018**	0.0001	0.90	16.44	0.001
DISTANCE FROM CENTER	**-0.0686**	0.009	0.87	-7.30	0.001
GOLDEN RING MALL	**1.5163**	0.645	0.98	2.35	0.05
EASTPOINT MALL	**1.6111**	0.648	0.97	2.49	0.05

Comparing Different Crimes Types

With or without special generators, a trip generation model is an ecological model that predicts crime origins and crime destinations. A point was made in Chapter 25 that these models are not behavioral, but are correlates of crimes. That is, the variables that end up predicting the number of crimes are not *reasons* (or explanations) for the crimes. Population almost always

enters the equation because, all other things being equal, zones with larger numbers of persons will have more crimes, both originating and ending in them. Similarly, low income status is frequently associated with high crime areas. It does not follow that low income persons will be more prone to commit crimes; it may be true but these models do not test that proposition (Ratcliffe, 2008). These are only correlates with crime in those environments. As was mentioned earlier, these variables are often correlated with many specific conditions that *may* be predictors of individual crime - poverty, drug use, substandard housing, and lack of job opportunities.

To see this, three separate models of specific crime types were run for robbery, burglary, and vehicle theft. For each crime type, the general model was tested for both the origin and the destination models. If a variable was not significant, it was dropped and the model was re-run.

Table 27.6:
Models for Specific Crime Types: Poisson-Gamma Origin Model

	All Crimes	Robbery	Burglary	Vehicle Theft
CONSTANT	3.483	1.1165	1.1165	-1.4994
POPULATION	0.0004	0.0004	0.0004	0.0005
INCOME EQUALITY	-0.0178	-	-	-0.0214
NON-RETAIL EMPLOYMENT	-0.0001	-0.0002	-0.0002	-0.0001
RETAIL EMPLOYMENT	-0.0002	-	-	-

Table 27.7:

Models for Specific Crime Types: Poisson-Gamma Destination Model

	All Crimes	Robbery	Burglary	Vehicle Theft
CONSTANT	4.5208	2.7489	0.7977	2.2903
POPULATION	0.0003	0.0003	0.0004	0.0004
INCOME EQUALITY	-0.0213	-0.0295	-0.0220	-0.0131
RETAIL EMPLOYMENT	0.0020	0.0019	-	0.0009
DISTANCE FROM CBD	-0.0714	-0.0965	-0.0365	-0.1013

The population variable appears in every single model. As mentioned, all other things being equal, the larger the number of persons in a zone, the more crime events will occur whether those events are crime productions (origins) or crime attractions (destinations). Similarly, relative income equality appears in four of the six crime-specific models with the coefficient always being negative. In general, zones with relatively lower incomes will have more robberies, burglaries, and vehicle thefts. The only model for which income equality did not appear was as an origin variable for burglaries; apparently, burglars come from zones with various income levels, at least in Baltimore.

The other general variables have more limited applicability. Retail employment predicts both total crime origins and total crime destinations, but only predicts specifically robbery destinations and vehicle theft destinations; the latter tend to occur more in commercial areas than not. On the other hand, non-retail employment appears to be important only as a crime origin variable; zones with less non-retail employment tend to produce more offender trips. Distance from the CBD only appears as a destination variable; the closer a zone is to the metropolitan center, the higher the number of crimes being attracted to that zone; this variable was not important in the origin model.

In other words, these models are measuring general conditions associated with crime, not causes *per se*. They capture the general contextual relationships associated with crime productions and attractions. But, they do not necessarily predict individual behavior. Nevertheless, the models can be used for prediction since the conditions appear to be quite general.

Adding External Trips to the Origin Model

After an origin and destination model has been developed, the next step is to add any crime trips that came from outside the modeling area (external trips). In this case, these would be trips that came from areas that were not in either Baltimore County or the City of Baltimore (the modeling area).

A simple estimate of external trips is obtained by taking the difference between the total number of crimes occurring in the study area (Baltimore County destinations) and the total number of crimes originating in the modeling area (Table 27.8).

The difference between the number of crime enumerated within Baltimore County and that originating from both Baltimore County and the City of Baltimore is 1,627. This is 3.9% of the total Baltimore County crimes. In general, it is important that the external trips be as small as possible. Ortuzar and Willumsen (2001) suggest that this percentage be no greater than 5% in order to minimize potential bias from not including those cases in the origin model. It is not an absolute percentage, but more like a rule of thumb; in theory, any external trips could bias the origin model. But, in practice, the error will be small if external crime trips are a small percentage of the total number enumerated in the destination county.

In this case, the condition holds. For the three types of crime modeled, the percentage of external trips was also less than 5%: robbery (4.0%), burglary (4.5%), and vehicle theft (1.4%). On the other hand, if the percentage of external trips is greater than approximately 5%, a user would be advised to widen the origin study area to include more zones in the model.

Predicting External Trips

If a model is being applied to another data set from which it was initially estimated, a problem emerges about how to estimate the number of external trips. It is one thing to apply simple arithmetic in order to determine how many trips originated outside the modeling area (as in Table 27.8). It is another to know how to calculate external trips when the model is being applied to other data. For the modeled zones, the coefficients are applied to the variables of the model (see 'Make Prediction' below). But, the external trips have to be estimated independently.

There is not a simple way to estimate external crime trips. Unlike regular trips that can be estimated through cordon counts, crime trips are not detectable while they are occurring (i.e., one cannot stand by a road and count offenders traveling by). Thus, they have to be estimated.

Table 27.8:
Estimating External Crime Trips in Baltimore County

Number of crimes ending in 325 Baltimore County zones:	41,969
Number of crimes originating in 532 Baltimore County/City zones:	40,342
Crimes from outside the modeling area:	1,627

Note: external trips are only added to the origin model since they are crime trips that originate outside the modeling area. They are not relevant for the destination model.

A simple method is to calculate the number of external trips for two time periods. For example, external trips could be calculated from a 2010 data set by subtracting the total number of crimes occurring in the modeling region from the total number of crimes occurring in the study area (e.g., as in Table 27.8 above). If a similar calculation was made for, say, 2012, then the difference (the 'trend') could be extrapolated. To take our example, between 1993 and 1996, there were 1,627 external trips. If the number of external trips turned out to be 1,850 for 1997-2000, then the difference (1,850 - 1,627 = 223) could be applied for future years. Essentially, a slope is being calculated and applied as a linear equation:

$$Y_i = 1850 + 223X_i \tag{27.32}$$

where Y_i is the number of crime origins during a four year period, I, and X_i is an integer for a four year period starting with the next period (i.e., the base year, 1997-2000, has integer value of 0). In other words, a linear trend is being extrapolated.

How realistic is this? For short time periods, linear extrapolation is probably as good a method as any. But for longer time periods, it can lead to spurious conclusions (e.g., crime trips from outside the region will always increase). Short of developing a sophisticated model that

relates crime trips to the growth of the metropolitan area and to other metropolitan areas within, say, 500 miles, a linear extrapolation is one of the few methods that one can apply.[10]

Make Prediction

In *CrimeStat*, external trips are added on the second page of the trip generation - Make prediction. This is a page where the modeled coefficients and any external trips are applied to a data set. There are two reasons why this is a separate page from the 'Calibrate model' page where the model was calibrated. First, the coefficients might be applied to another data than that from which it was calibrated. For example, one might calibrate the model with a data set from 2008-2010 and then apply to a data set covering 2011-2013. Similarly, one might take future year forecasts (e.g., 2025) and apply the model. In effect, the model would be predicting the number of future crimes *if* the same conditions hold over the time frame.

A second reason for separating the calibration and application pages is to add external trips to the origin zones. As mentioned above, external trips are, by definition, those that were not modeled in the calibration. They have to be calculated independently of the model and then added to the estimates.

Thus, the 'Make prediction' page allows these operations to occur. Figure 27.10 shows the page. There are several steps that have to be implemented for this page to be operative.

1. The data file has to be input as either the primary or secondary file (not shown in the image). In this example, the same data set is being used as was used for the calibration. But, if it is a different data set, that will need to be input in the Data Setup section. Whether the input data set is a primary file (the usual occurrence) or a secondary file needs to be specified. Also, indicate whether the applied model is to be an origin or destination model. In Figure 27.10, it is defined as an origin file.

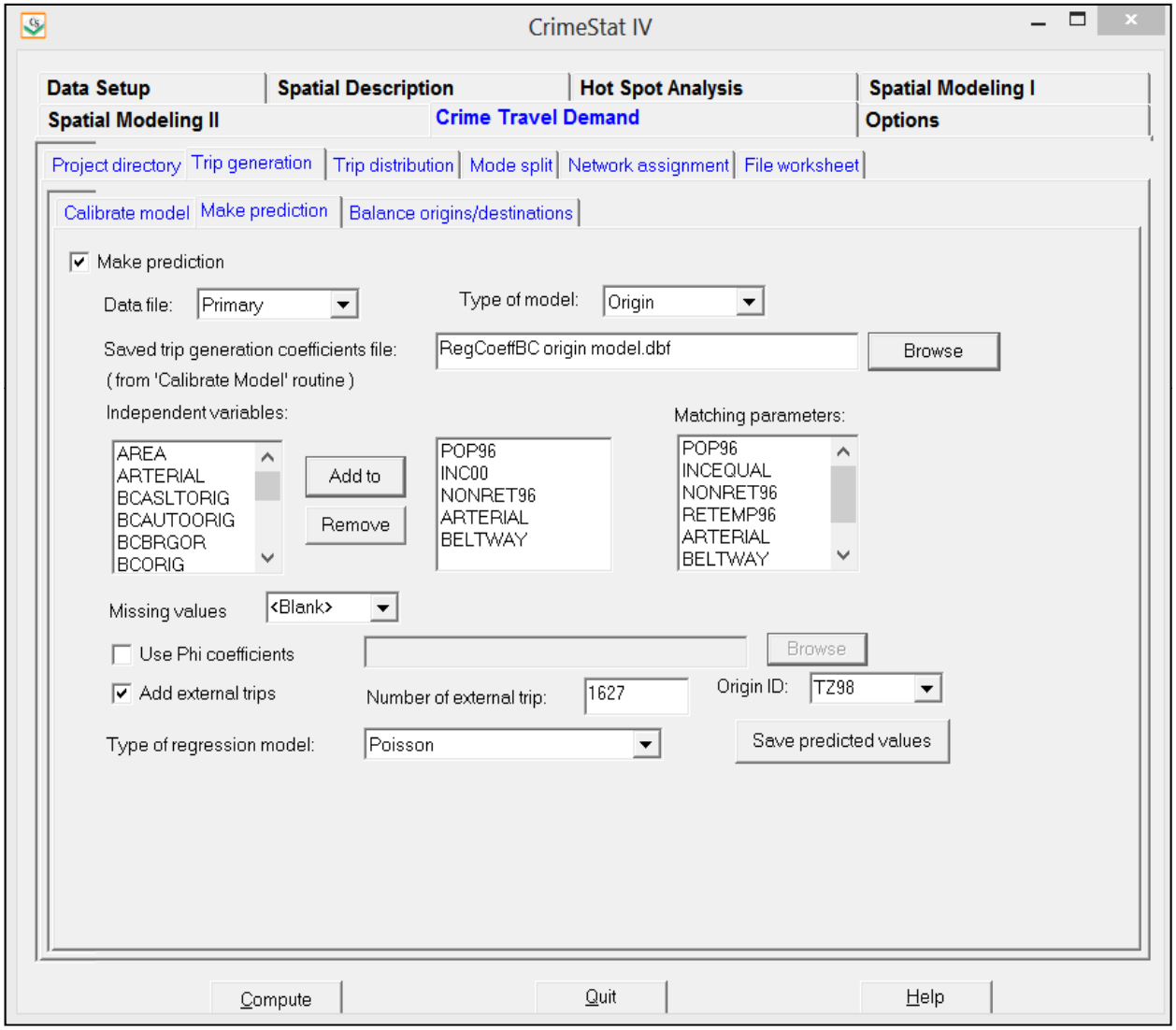
2. A trip generation coefficients file needs to be input. These were the estimated coefficients from the calibration stage. Inputting this file brings in the coefficients in the order in which they were saved. They are listed in the 'Matching parameters' dialogue box on the right side of the page.

---

[10] An alternative might be to use cordon counts from major highways coming into the region and assume that crime trips represent a constant proportion of those trips. Thus, if the total number of estimated external highway trips increases by 5%, one could assume that the external trips also increase by 5%. While this is plausible, it is not necessarily an accurate estimate. Talk to your Metropolitan Planning Organization or the State Department of Transportation if you are interested in developing this type of model as you will need their estimates of external trips.

Figure 27.10:

"Make Prediction" Setup Page



3. On the left side of the page are listed all the variables in the input data set (primary or secondary file). In the middle box, the variables are added in the **same order** as in the matching parameters box. That is, each independent variable needs to be matched to the variable from the coefficients file, one for one. **This is very important**. The names do not have to be the same (e.g., if the model was calibrated with data set and applied to another, the variable names may not be identical). But the content and order of the variables needs to be the same. In the example, the first variable in the coefficients file is INCEQUAL. The selected variable in the middle box has to be the income equality variable (whatever its name). In the example, the same data set is being used so the names are identical. This is repeated for each of the independent variables in the coefficients file.

4. Next, any missing value codes are specified in the missing values box. Any records with a missing value for *any* of the selected independent variables will be dropped from the calculation. In the example, there are no missing value codes applied other than the default blank field.

5. If external trips are to be added, the external trips box must be checked. External trips could be applied in an origin model, but not in a destination model. If they are to be added, the number of trips should be specified in the 'Number of external trips' box and the zone ID field for the file indicated; in the example, 1627 is added as external trips and the TAZ field is specified as the ID variable (TZ98).

6. The type of model to be applied is indicated in the "Type of regression model" box. There are only two choices: Poisson (the default) and Normal (OLS). Since the coefficients are being applied to the data, no over-dispersion correction is necessary (since it was probably used in calibrating the model).

7. Finally, the output file name is defined in the 'Save predicted values' box.

For each zone, the routine will then take the appropriate variable from the input data set and apply the matching coefficient from trip generation coefficients file to produce a predicted estimate of the number of trips. To calculate this value, for the OLS model, the routine will use equation 27.2 above while for the Poisson model, the routine will use equation 27.6 above. For the latter, it will then raise the predicted log value to the power, *e*, to produce a prediction for the expected number of crime trips:

$$\lambda_i = e^{Ln(\lambda_i)} \tag{27.33}$$

If external trips are added, a new zone is created called EXTERNAL in the ID field that was indicated on the page. Then, the specified number of external trips is simply placed in that field with zeros being placed for the values of all the remaining variables in the file. By default, the output name for the predicted number of crimes will be called PREDORIG for an origin model and PREDDEST for a destination model. An example data set is available on the *CrimeStat* download page.

> Note: for a destination model, this 'Make prediction' operation is not necessarily needed if the same data set is used for calibration and prediction. This step is primarily for the origin file

Balancing Predicted Origins and Destinations

After the origin model and destination model are calibrated and applied to a data set, the final step in trip generation is to ensure that the number of predicted origins equals the number of predicted destinations. This is necessary for the next stage of crime travel demand modeling - trip distribution. Since a trip has both an origin and a destination, the total number of origins *must* equal the total number of destinations. This is an *absolute* requirement for the trip distribution model to work. The routine will return an error message if the number of origins does not equal the number of destinations.

If the Poisson model is used for calibration, the routine ensures that the number of predicted trips equals the number of input trips. Further, if the calculation of external trips has been obtained by subtracting the total number of predicted origins from the total number of predicted destinations, and if the external trips are then added to the predicted origins, then most likely the total number of origins will equal the total number of destinations. However, because of rounding-off errors and inconsistent external trip estimates, it is possible that the sums are not equal.

Consequently, it is important to balance the predicted origins and destinations to ensure that no problems will occur in the trip distribution model. There are two ways to do this in *CrimeStat*. First, the number of predicted destinations is held constant and the number of predicted origins is adjusted to match this number. This is the default choice. Second, the number of predicted origins is held constant and the number of predicted destinations is adjusted to match this number.

The calculation is essentially a multiplier that is applied to each zone. If destinations are to be held constant, the multiplier is defined as the ratio of total destinations to total origins:

$$M_j = \frac{Sum\ of\ crimes\ by\ destinations}{Sum\ of\ crimes\ by\ origins} = \frac{\sum_{j=1}^{N} X_j}{\sum_{i=1}^{M} X_i} \quad (27.34)$$

The predicted number of origins is multiplied by M_j. If, on the other hand, the origins are to be held constant, the multiplier is defined as the ratio of total origins to total destinations:

$$M_j = \frac{Sum\ of\ crimes\ by\ origins}{Sum\ of\ crimes\ by\ destinations} = \frac{\sum_{i=1}^{M} X_i}{\sum_{j=1}^{N} X_j} \quad (27.35)$$

The predicted number of destinations is multiplied by M_i. The multiplication simply ensures that the sums of the predicted origins and predicted destinations are equal.

The third page in the trip generation model is the 'Balance predicted origins & destinations' page. Figure 27.11 shows the setup for this page. The steps are as follows:

1. The box is checked indicating that it is a balancing operation.

2. The predicted origin file is input and the predicted origin variable is identified. In the example, the predicted origin file is called 'PredictedOrigins.dbf' and the field with the predicted numbers was called PREDORIG.

3. The predicted destination file is input and the predicted destination variable is identified. In the example, the predicted destination file is called 'PredictedDestinations.dbf' and the field with the predicted numbers was called PREDDEST.

Note that these files are input on this page and not on the primary or secondary file pages.

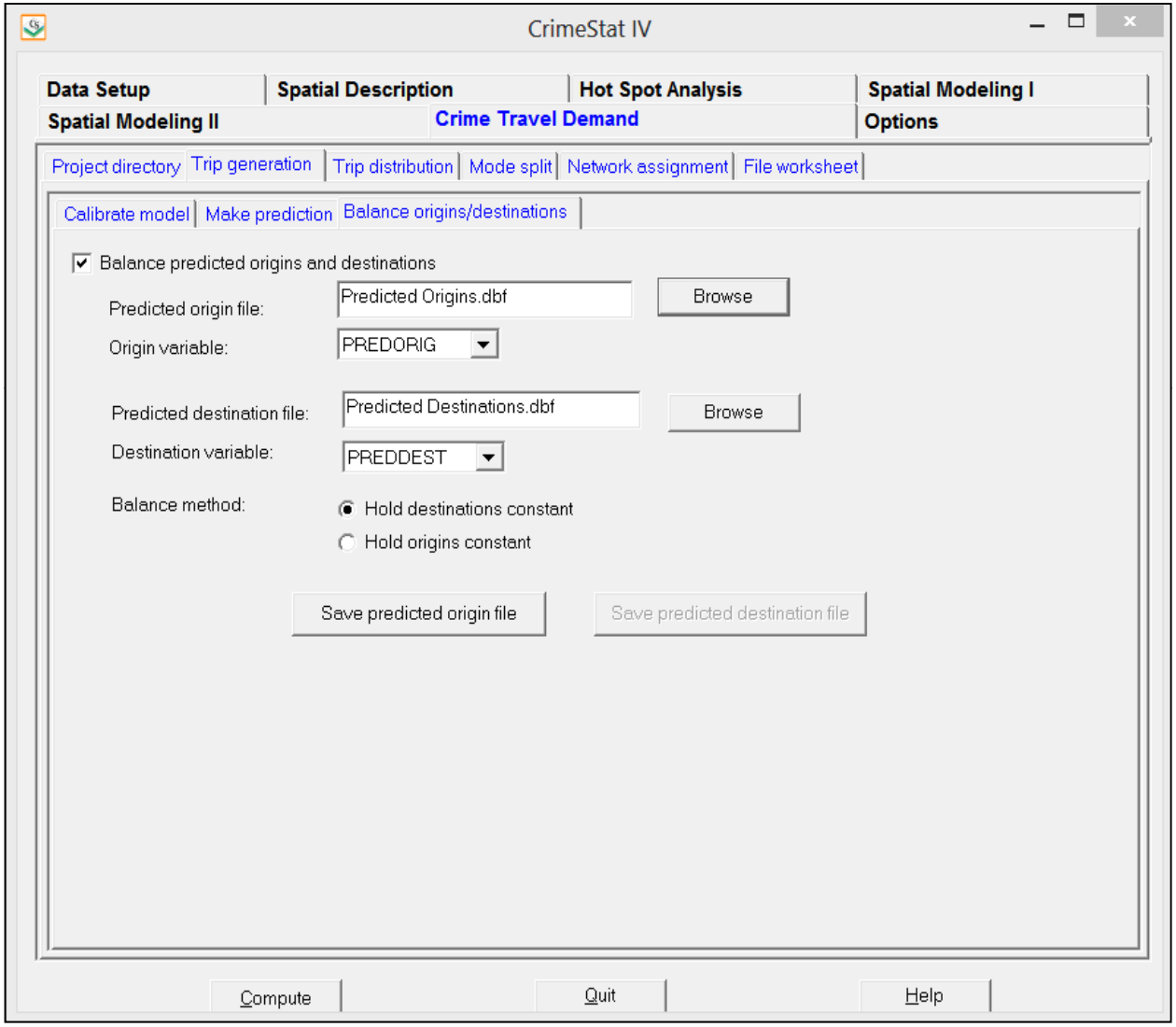
4. Next, the type of balancing is specified - Holding destinations constant (the default) or holding origins constant. In the example, the destinations are to be held constant.

5. Finally, the output file is specified. If the origins are to be adjusted, then only the origin file is saved. If the destinations are to be adjusted, then only the destination file is saved. In other words, the adjustment is applied to only one of the two predicted crime files. In the example, the file was named 'AdjustedPredictedOrigins.dbf' (not shown) since the origin file was adjusted.

The output produces a new column with the adjusted values. Table 27.9 shows the origin output for the Baltimore data of the first 11 records. Once the balancing has been completed, the trip generation model is finished and the user can go on to the trip distribution model. In other

Figure 27.11:

Balance Predicted Origins and Destinations Setup



words, the output file ensures that both the predicted origin file (crime productions) and predicted destination file (crime attractions) are balanced.

Table 27.9:
Adjusted Data Should Have These Fields

Zone	PREDICTED	ADJORIGIN
0001	225.818482	225.850955
0002	187.527819	187.554785
0003	320.877458	320.923600
0004	75.096631	75.107430
0005	44.981775	44.988243
0006	32.574758	32.579442
0007	107.334835	107.350270
0008	74.683931	74.694671
0009	76.425236	76.436226
0010	34.183846	34.188762
0011	66.975803	66.985434
etc	etc	etc

Strengths and Weaknesses of Regression Modeling of Trips

As mentioned earlier, the use of regression for producing the trip generation model has its strengths and weaknesses. The advantages are that, first, the approach is applicable to crime incidents. Unlike regular travel behavior, crime trips have to be inferred from police reports. Thus, starting with counts of the number of crimes occurring in each zone and the number of crimes that originate from each zone, a model can be constructed.

Second, the use of a non-linear model, such as the Poisson, allows more complex fitting of crime counts. In the early 1970s when trip generation models were starting to be implemented in Metropolitan Planning Organizations around the U.S., the major type of regression modeling available was OLS. At that time, researchers could not demonstrate that this method was reliable in terms of predicting travel. However, with the availability of software for conducting Poisson and other non-linear models, that criticism is no longer applicable. The Poisson model is very well behaved with respect to count data. It does not produce negative estimates. It requires high levels of an independent variable to produce a slight effect in the dependent variable, but that the level increases as the values of the independent variable increase. It maintains constancy between the sum of the input counts and the sum of the predicted counts. Non-linear models are much more realistic for modeling trips than OLS.

Third, the use of a multivariate regression model allows multiple variables to be included. In our example, there were four independent variables in the reduced origin and destination models. Trip tables, on the other hand, typically only have three or four independent predictors; it becomes too complicated to keep track of multiple conditions of predictor variables. Thus, a more complex and sophisticated model can be produced with a regression framework.

Fourth, and finally, a regression framework allows for complex interactions to be estimated. For example, the log of an independent variable can be defined. An interaction between two of the independent variables can be examined (e.g., median household income for those zones having a sizeable amount of retail employment). In the trip table approach, these interactions are implicit in the cell means. Thus, overall, the regression framework allows for a more complex model than is available with a trip table approach.

On the other hand, there are potential problems associated with a regression framework. First, the regression coefficients can be influenced by zone size. Since the model is estimating differences between zones (i.e., differences in the number of crimes as a function of differences in the values of the independent variables), zone size affects the level of those differences. With small zone sizes, there will be substantial differences between zones in both the independent and dependent variables. Conversely, large zone sizes will minimize within-zone differences, but will usually increase the estimate of the between-zone differences. The result could be an exaggeration of the effect of a variable that would not be seen with small zone geography. As was argued in Chapter 25, one should choose the smallest zone geography that is practical in order to minimize this problem.

Second, a point that has been repeated again and again, these models are not behavioral explanations. They represent ecological correlations with crime trips. It is important to not try to convert these models into explanations of offender behavior. Too often, researchers have jumped to conclusions about individuals based on the relationships with environments and neighborhoods. It is important to not do this. This criticism, incidentally, applies both to the trip table as well as the regression approach to trip generation modeling.

The new generation of travel demand models are specifically behavioral and involve modeling the behavior of specific individuals. Probabilities are calculated based on individual choice and a micro-simulation routine can apply these probabilities to a large metropolitan area (Shifton et al, 2003; Recker, 2000). While this approach offers some definite theoretical advantages and is the subject of much current research, to date there has not been a demonstration that this approach is more accurate at predicting trips than the tradition trip-based travel demand model. For crime, such an approach would have to be simulated.

Conclusion

In summary, the trip generation model is a valuable tool for predicting the number of crimes that originate in each zone and the number of crimes that end in each zone. Even if the model is not behavioral, the model can be stable and useful for many years in the future. It is best thought of as a *proxy model* in which the variables in the models are proxies for conditions that are generating crimes, either in terms of environments that produce offenders or in terms of locations that attract them.

In the next chapter, we will examine the second stage in the travel demand model - trip distribution. In that stage, the predicted crime origins and the predicted crime destinations are linked to produce crime trips.

References

Boswell, M. T. & Patil, G. P. (1970). "Chance mechanisms generating negative binomial distributions". In *Random Counts in Scientific Work*, Vol. 1, G. P. Patil, ed., Pennsylvania State University Press:University Park, PA, 3-22.

Bowers, K. & Hirschfield, A. (1999). Exploring links between crime and disadvantage in North-West England: An analysis using Geographic Information Systems. *International Journal of Geogrpahical Information Science*, 13,b 159-184.

Bursik, R. J., Jr. & Grasmick, H. G. (1993). Economic deprivation and neighborhood crime rates, 1960-1980. *Law and Society Review*, 27, 263-268.

Cameron, A. C. & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press: Cambridge, U.K.

Cameron, A. C. & Windmeijer, F. A. G. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, 14(2), 209-20.

Chiricos, T. (1987). Rates of Crime and Unemployment *Social Problems*, 34, 187-211

Cohen, L.E. & Felson, M. (1979) Social change and crime rate trends: a routine activity approach, *American Sociological Review*, 44: 588-608.

Culp, M. & Lee, E. J. (2005). Improving travel models through peer review. *Public Roads*, 68 (6), FHWA-HRT-05-005. Federal Highway Administration, U.S. Department of Transportation: Washington, DC. http://www.fhwa.dot.gov/publications/publicroads/05may/07.cfm. Accessed April 28, 2012.

Der, G. & Everitt, B. S. (2002). *A Handbook of Statistical Analyses using SAS*. Chapman & Hall/CRC: London.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis, Second Edition*. John Wiley & Sons: New York.

Ehrlich, I. (1975). On the relation between education and crime. In F. T. Juster (ed), *Education, Youth and Human Behavior*. McGraw-Hill: New York, 313-337.

Fotheringham, A. S., Brunsdon, C. & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons: New York.

References (continued)

Fowles, R. & Merva, M.. (1996). Wage Inequality and Criminal Activity, *Criminology*, 34, 163-82.

Freedman, David A. (1999). Ecological inference and ecological fallacy. *International Encyclopedia of the Social and Behavioral Sciences*, Technical Report No. 549, October. http://www.stanford.edu/class/ed260/freedman549.pdf. Accessed March 26, 2012.

Hagan, J. & Peterson, R. (1994). *Inequality and Crime*. Stanford University Press: Palo Alto, CA.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56, 1030-1039.

Hensher, D. A. & Button, K. J. (2002). *Handbook of Transport Modeling*. Elsevier Science: Cambridge, UK.

ITE (2003). *Trip Generation* (7th edition). Institute of Transportation Engineers: Washington, DC.

Kohfeld, C. W. & Sprague, J. (1988). Urban unemployment drives crime. *Urban Affairs Quarterly*, 24, 215-241.

Langbein, L. I. & Lichtman, A. J. (1978). *Ecological Inference*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-010. Beverly Hills and London: Sage Publications.

Messner, S. (1986). Economic inequality and levels of urban homicide, *Criminology*, 23, 297-317.

Miaou, S.P (1996). *Measuring the Goodness-of-Fit of Accident Prediction Models*. FHWA-RD-96-040. Federal Highway Administration, U.S. Department of Transportation: Washington, DC.

Microsoft (2012). SKEW - skewness function, *Microsoft Office Excel 2010, Microsoft: Redmond, WA. http://office.microsoft.com/en-us/excel-help/skew-HP005209261.aspx. Accessed May 21, 2012.*

Nannen, V. (2003). *The Paradox of Overfitting*. Artificial Intelligence, Rijksuniversitat: Groningen, Netherlands. http://volker.nannen.com/pdf/the_paradox_of_overfitting.pdf. Accessed March 11, 2010.

References (continued)

NCHRP (1998). *Integration of Land Use Planning with Multimodal Transportation Planning*. Project 8-32(3). Prepared by Parsons Brinkerhoff Quade & Douglas, Inc. for the National Cooperative Highway Research Program, Transportation Research Board, National Research Council: Washington DC. October.

NIST (2004). Gallery of distributions. *Engineering Statistics Handbook*. National Institute of Standards and Technology: Washington, DC. http://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm. Accessed May 21, 2012.

Newman, O. (1972). *Defensible Space: Crime Prevention Through Urban Design*. Macmillan: New York.

Ortuzar, J. D. & Willumsen, L. G. (2001). *Modeling Transport* (3rd edition). J. Wiley & Sons: New York.

Park, R. & Burgess, E. (1924). *Introduction to the Science of Sociology*. Chicago University Press: Chicago.

Pribyl, O & Goulias, K. G. (2005). Simulation of **d**aily **a**ctivity **p**atterns **i**ncorporating **i**nteractions **w**ithin **h**ouseholds: Algorithm overview and performance. *Transportation Research Record*, 1926 (January), 135-141. http://trb.metapress.com/content/r7u36h005758h304/**.** Accessed May 9, 2012.

Radford, N. (2006). The problem of overfitting with maximum likliehood . CSC 411: Machine Learning and Data Mining, University of Toronto: Toronto, CA. http://www.cs.utoronto.ca/~radford/csc411.F06/10-nn-early-nup.pdf Accessed March 11, 2010.

Ratcliffe, J.H. (2008). The magnitude of the crime challenge (Chapter 3). *Intelligence-Led Policing*, Willan Publishing: Cullompton.

Recker, W. (2000). A bridge between travel demand modeling and activity-based travel analysis. *Center for Activity Systems Analysis*. Paper UCI-ITS-AS-WP-00-11. http://repositories.cdlib.org/itsirvine/casa/UCI-ITS-AS-WP-00-11/. Accessed May 23, 2012.

Shaw, C. R. & McKay, H. D. (1942). *Juvenile Delinquency in Urban Areas*. Chicago: University of Chicago Press.

References (continued)

Shifton, Y., Ben-Akiva, M., Proussaloglu, K., de Jong, G., Popuri, Y., Kasturirangan, K., & Bekhor, S. (2003). Activity-based modeling as a tool for better understanding travel behaviour. *Conference Proceedings*. 10th International Conference on Travel Behaviour Research, Lucerne, Switzerland. August. http://www.ivt.ethz.ch/news/archive/20030810_IATBR/shiftan.pdf. Accessed May 23, 2012.

Shoup, D. (2002). Roughly right vs. precisely wrong. *Access*, No. 20, Spring. 20-25.

Stack, S. (1984). Income inequality and property crime, *Criminology*, 22, 229-257.
Thrasher, F. M. (1927). *The Gang*, University of Chicago Press: Chicago.

Venables, W.N. & Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus (second edition)*. Springer-Verlag: New York.

Wilson, J.Q. & Kelling, G. (1982) Broken Windows: The Police and Neighborhood Safety. *Atlantic Monthly*, March. 29-38.